Journal of Computer Science

Review

Artificial Intelligence Techniques and External Factors used in Crime Forecasting in Violence and Property: A Review

[1]Rebaz Mala Nabi, [2]Soran Ab. M. Saeed and [3]Habibollah Haron

[1]Computer Network, Sulaimani Polytechnic University, Sulaimani, KRG, Iraq
[2]Sulaimani Polytechnic University, Sulaimani, KRG, Iraq
[3]Qaiwan International University, Sulaiman, KRG, Iraq

Article History
Received: 22-11-2019
Revised: 06-01-2020
Accepted: 15-02-2020

Corresponding Authors:
Rebaz Mala Nabi
Computer Network, Sulaimani Polytechnic University,
Sulaimani, KRG, Iraq
Email: rebaz.nabi@spu.edu.iq

Abstract: Crime forecasting is beneficial in providing useful information to authorities in planning effective crime prevention measures. The two types of analysis used in crime forecasting are univariate and multivariate. Comparatively, multivariate analysis provides better forecasting accuracy because of its ability to discover crime patterns not previously seen. Crime is strongly influenced by several external factors, including economic, social and demographic. Hence, an analysis is needed to identify and select relevant factors that influence crime and can later be used to improve forecasting accuracy. Neighborhood Component Analysis (NCA) is a reliable form of analysis for identifying significant relationships between factors and crime data. Several model types have been introduced in crime forecasting, including statistical and artificial intelligence models. Recently, the artificial intelligence model has come into favour because of its ability to handle nonlinearity patterns in crime data well. Within the artificial intelligence model, Gradient Tree Boosting (GTB) shows good performance as it produces a robust and reliable forecast result. GTB uses least square function as a loss function for error fitting during training. Findings show that, in addition to using least square function, implementing other standard mathematical functions that fit to the crime data increases forecasting accuracy. In other cases, both NCA and GTB are sensitive to parameters input. Dragonfly Algorithm (DA) is a promising, nature inspired metaheuristic algorithm that is capable of solving such problems.

Keywords: Crime Forecasting, Crime, GTB, DA, NCA, External Factors

Introduction

Forecasting is a means of predicting or making a statement regarding uncertain (future) events, based on past or present knowledge under specific criteria. It is widely applied in estimating the degree of risk or uncertainty in many areas, including but not limited to geographical, financial, economic, engineering, security, health and many interdisciplinary areas (Jain and Kumar, 2007). Because provided data plays an important role in forecasting, the forecasting model itself is developed based on that data.

As we know, the data in a real-world problem is noisy and unstable. Thus, the greatest challenge in creating the forecasting model is how well it is able to handle such data. Time series, cross-sectional and longitudinal data are all used in forecasting.

The majority of forecasting models are based on time series data. This is because the future values of a physical variable, which are measured in time, at discrete intervals or on a continuous basis, are needed in important planning, design and management processes (Jain and Kumar, 2007). Time series data constitute a set of observations with a discrete, time stochastic nature (De Oliveira and Ludermir, 2014). It consists of historical observation of past data within the same variable, which can subsequently be used to develop a time series model describing the underlying relationship. The developed model is then used to extrapolate the time series into the future (Alwee *et al*., 2013). Time series models are useful when limited knowledge exists regarding the underlying data generating process and when there is an absence of a satisfactory explanatory model relating the estimation variables to other explanatory variables (Khashei and Bijari, 2011). As previously mentioned, time series models have been widely applied in many areas (Khashei and Bijari, 2011; Chen and Tanuwijaya, 2011; Zhang, 2003).

The two types of time series model, namely linear and non-linear, are based on time series data structure and nature. Each time series model is developed to capture the linearity or non-linearity of the time series data. Linear time series models are primarily based on statistical models, such as linear regression, exponential smoothing, Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA). The linear time series model is a parametric technique which assumes that the obtained time series data are stationary and linear (Alwee *et al*., 2013). In other words, forecasting requires that the time series data be first reconstructed as stationary and linear. Although the linear time series model has proven to be a good forecasting model in past decades, because it cannot capture non-linear data, its usefulness is limited. Several non-linear time series models have been introduced by researchers to overcome the difficulty in capturing non-linear time series data (Zhang, 2003; Cheng *et al*., 2008; Yeh, 2013; Khashei and Bijari, 2011). For the most part, the non-linear time series model is non-parametric and based on artificial intelligence techniques, such as Artificial Neural Network (ANN), Support Vector Regression (SVR), Evolutionary Algorithm (EA) and swarm intelligence. In recent years, artificial intelligence techniques have been extensively used in time series models. This has come about because artificial intelligence techniques contain some non-linear functions which are able to detect non-linear patterns in data, thus improving forecasting performance (Han and Wang, 2009).

In time series models, there are two types of analysis: univariate and multivariate. The term, time series model, describes a precise objective achieved by both analyses. Univariate analysis is the simplest form of analysis, only requiring one variable of time series data to develop the forecasting model. It forecasts future outcome with an assumption that the both past and future behaviour of the data are similar. The main purpose of univariate analysis is to describe the pattern of behaviour of the data. Multivariate analysis uses data from more than one time series in model development. The analysis finds cross-correlation among data from multiple time series (Khashei and Bijari, 2011). The purpose of multivariate analysis is to find the relationship between dependent and independent variables. It is most useful when discovering a new pattern of data that has not occurred in the past (Alwee *et al*., 2013). Multivariate analysis provides better forecasting accuracy compared to univariate analysis (Han and Wang, 2009). Thus, in this research, only the multivariate time series analysis will be addressed.

Over the last decade, forecasting research interest has shifted from the statistical model to the artificial intelligence model. One of the reasons for this is that the statistical model is incapable of handling abrupt environmental or social changes (Baliyan *et al*., 2015). Thus, it will negatively affect the performance accuracy of the forecasting model. In addition, in a real-world problem, the time series data contain a mixture of both linear and non-linear patterns. As previously mentioned, the statistical model captures the linear pattern but not the non-linear. With the recent growth in technology and greater knowledge, the application of artificial intelligence techniques in forecasting promises further enhancement and improvement in forecasting performance accuracy. In addition, artificial intelligence models have the capability to global search, which make them more efficient in handling complex environments (Baliyan *et al*., 2015).

In addressing the artificial intelligence model several issues arise. For example, Artificial Neural Network (ANN) has been widely used in forecasting, but it suffers from a lack of parameters control, possibility of overfitting and network weights uncertainty (Alwee *et al*., 2013). According to (Khashei and Bijari, 2011), ANN performs worse in some specific cases than the existing statistical model when the data is linear and without much disturbance. In other words, ANN might not perform better when handling a linear relationship. Thus, it is unwise to apply ANN blindly to any type of data (Khashei and Bijari 2011). As another example, Support Vector Regression (SVR) suffers a parameter optimization problem such that a highly effective model can only be built with carefully selected SVR parameters. In other words, SVR is sensitive to parameter determination. This is because kernel-parameters are the few adjusted parameters in SVR which control the complexity of the resulting hypothesis (Wu *et al*., 2009; Amroune *et al*., 2018). In addition, the main difficulty in SVR lies in selecting the best available kernel function, corresponding to the feature space and the parameter of this kernel function (Foucquier *et al*., 2013).

Most artificial intelligence models are sensitive to parameters. Such problems have been addressed by several researchers, especially in ANN (Yeh, 2013; Rather *et al*., 2017; Hipp and Yates, 2011) and SVR (Wu *et al*., 2007; Wu *et al*., 2009; De Oliveira and Ludermir, 2014). These researchers have applied a promising solution to such problems by integrating other artificial intelligence techniques into existing artificial intelligence models, developing a new hybrid model (Wu *et al*., 2009; De Oliveira and Ludermir, 2014; Hipp and Yates, 2011). Such approaches have proven to be successful because the hybrid model, which is more robust to changes in time series patterns, is able to outperform other singular models (Cook and Durrance, 2013).

Introduction to Crime Forecasting

In the real world, crime is a part of society which is unpredictable by police (Bye, 2007). Crime rate statistics demonstrate the degree of lack of public safety within the country. They provide information useful to governments and police in planning crime prevention

measures. The term crime analysis refers to a discipline practiced by the policing community (Bye, 2007). The analysis of crime data may help in the understanding of behavioural trends and future values may be forecasted from past observations (Song *et al*., 2018; Xiao *et al*., 2018). Thus, crime forecasting represents a promising solution which affects the relative well-being of people's life and properties. Police will generally take a common approach in handling crime such as intuition, experience, evidence and collected information from witnesses. Such approaches are impractical because the police only take an action after a crime has already occurred. Hence, the role of crime forecasting will be to enable early crime prevention measures (Hapfelmeier and Ulm, 2013). The advantages of crime forecasting include the prevention of recurrent crimes in specific areas or regions through analysis of the pattern of past crime occurrence, help in appropriate resource allocation within a community, allowing for better police coverage and the provision of useful information to authorities for the planning of efficient solutions in crime prevention measures.

Although crime forecasting has proven promising in estimating future crime rates, it is still rarely applied in most countries, including Malaysia (Hapfelmeier and Ulm, 2013). This is due to several challenges inherent in crime data and model accuracy. In the real world, historical crime data is difficult to obtain and therefore limited in availability (Alwee *et al*., 2013). Most countries including Malaysia treat such data as confidential. In addition, the available data, itself, is noisy and may be incomplete. Due to insufficient data, it is therefore more difficult to develop a crime model. To make matters worse, crime model accuracy performance is heavily affected by inconsistencies in crime data structure. In criminology fields, researchers attempt to forecast crime based on two objectives. These are by the identification of potential crime hotspots in a specific area or region and the identification of future crime patterns. Different researchers use different methods, which depend on data type and nature. In crime forecasting, the most used crime data representation is time series data. This type of crime data is extensively used by researchers to forecast crime rates. Time series data are collections of historical data from specific time ranges which can later be used to develop a time series model. In the literature, several types of crime forecasting have been introduced, including the statistical model, the artificial intelligence model and a mixture of both aforementioned models, which can be called a hybrid model.

The statistical model adapts several statistical techniques to analyse past or present crime data trends to predict future outcomes or events. In most work, the statistical model remains the most widely used and it is considered the conventional method in crime modelling. Foucquier *et al*. (2013) applied the Autoregressive (AR) model to forecast crime rate in an urban area located in Chicago, USA. The proposed model was able to forecast crime numbers with an accuracy of 84% projecting one-year-ahead and 80% looking ahead two years. Cesario *et al*. (2016) used the ARIMA model to forecast property crime in China. The results show that the ARIMA model easily outperforms the standard statistical model when it fits the data well and thus, further enhances forecasting accuracy. Bye (2007) implemented the ARIMA model to analyse the effects of several external factors, such as alcohol consumption, in violent crime in Norway. This model has proven to perform well in determining the relationship between time series crime data and external factors, offering an explanation for the changes in violent crime rates over time. Chen *et al*. (2008) used linear regression, adaptive regression and decision stump algorithms to evaluate violent crime patterns within a crime dataset. The results indicate that the model is very effective in forecasting crime data based on given input. Mcclendon and Meghanathan (2015) implemented three statistical methods, namely random walk, Brown's simple exponential smoothing and Holt's two-parameter linear exponential smoothing on univariate time series crime data in Pittsburgh, USA. The experiment results show that Holt's two-parameter linear exponential smoothing performs better compared to the other methods in precinct-level univariate time series crime data. Hapfelmeier and Ulm (2013) applied short term crime forecasting to evaluate crime patterns in Kedah, Malaysia. This model used exploratory data analysis and moving average methods to forecast future crime patterns in univariate time series data.

Artificial Intelligence Techniques Involved in Crime Forecasting

The artificial intelligence-based forecasting model adopts a soft computing, or machine learning, technique. This approach has gained popularity recently because of its ability to enhance forecasting performance relative to the statistical model. This study focuses on four artificial intelligence techniques applied in forecasting models. These are Artificial Neural Network (ANN), Support Vector Regression (SVR), Random Forest (RF) and Gradient Tree Boosting (GTB).

Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is one of the artificial intelligence techniques introduced in 1943 by Warren McCulloh and Walter Pitts. Based on the human central nervous system, it is capable of evaluating functions that rely on a large number of inputs which are generally unknown or depend on other functions' output. ANN is widely applied in the computer science field because of its capabilities in machine learning and

pattern recognition. It is a parallel distributed information processing structure which consists of processing elements that can be considered as neurons (Arora *et al*., 2019; Cook and Durrance, 2013; Xiao *et al*., 2018).

Neurons in ANN can process a local memory and perform localized information processing operations. The neurons are interconnected with complex, non-linear, unidirectional signal channels, referred to as connections, that later form a massive parallel network. Each neuron has only a single output connection but can have more than one collateral connection branch. Each branch carries the neuron's output signal, which constitutes a type of mathematical form. The neuron performs its own process, based both on the input received via its connection channel and on values stored in its local memory. This process does not relate to other neuronal processes.

Support Vector Regression (SVR)

Support Vector Regression (SVR) is a machine learning, or artificial intelligence, technique introduced by Vladimir Naumovich Vapnik and Alexey Yakovlevich Chervonenkis in 1963. It adapts the structural risk minimization inductive principle to produce a good generalization on a restricted number of learning patterns (Wu and Lu, 2012). SVR is derived from the computation of a linear regression function in high dimensional feature space in which input data are mapped via non-linear functions (Chandrasekar *et al*., 2015). Its purpose is to minimize the upper boundary of the generalization error to achieve generalized performance. Thus, SVR is capable of generalizing the unseen data (Chandrasekar *et al*., 2015). SVR has been applied in various areas including time series forecasting.

Based on SVR, as explained by (Wu and Lu, 2012), it is assumed that there is a given training set of $(x_i, y_i)i = 1l$, where $x \in R^d$ is the *ith* input vector, $y_i \in R^d$ is the *i*th forecast output of y_i, d is the embedding dimension of the time series and l is the number of training data. The objective of SVR is to find the best function from possible function sets:

$$\left\{ f \mid f(x) = w^T x + b, w \in R^d, b \in R \right\} \tag{1}$$

where, w^T is the estimated weight factor, which is obtained from the minimized regularized risk function and b is a threshold. The mentioned regularized risk functions are shown as follows:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} L\left(y_i, f\left(x_i\right)\right) \tag{2}$$

To find the best function f, the regularized risk is minimized where $C>0$ is a regularization factor, $\|.\|$ is 2-norm and L (.,.) is a loss function. Once found, SVR sparsity is induced via application of the q-insensitive loss function to produce the q-tube. This will allow a portion of the forecasting to fall within the boundaries of the produced q-tube. The mentioned q-insensitive loss function is defined as follows:

$$L\left(y, f(x)\right) = \begin{cases} 0, & \left| f(x) - y \right| < q \\ \left| f(x) - y \right| - q, & Otherwise \end{cases} \tag{3}$$

From Equation 3, SVR is formulated as the minimization based on the following definition of $1/2||w||^2+CL(\xi_i.\xi_i*)l_i = 1$ and is subject to:

$$\begin{cases} w^T x_i + b - y_i \le q + \xi_i \\ y_i - w^T x_i - b \le q + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 0, ..., l \end{cases} \tag{4}$$

where, ξ_i and ξ_i^* are slack variables used to calculate the errors values of the up and down sides outside the boundaries that have been determined by the q-tube respectively. Finally, SVR will perform nonlinear mappings into higher dimensionality space using regression function defined as follows:

$$f(x) = \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) K\left(x_i, x_j\right) + b \tag{5}$$

where, α and α^* are multipliers, Lagrange and $K(x_i, x_j)$ is a kernel function. In machine learning theories, the most widely used kernel functions are linear kernel, polynomial kernel and Gaussian kernel (Wu and Lu, 2012). Each of the kernel functions is defined as follows:

$$Linear\ Kernel\ x_i^T . x_j \tag{6}$$

$$Polynormal\ Kernel\left(x_i . x_j + g\right)^z \tag{7}$$

$$Gaussian\ Kernel \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{8}$$

where, parameter g is an influence of higher-order versus lower-order terms in the polynomial, parameter z is a degree of polynomial kernel function and parameter σ is an epsilon-insensitive value in Gaussian kernel. Kernel functions play major roles in SVR accuracy performance.

Random Forest (RF)

Random forest is an artificial intelligence technique based on aggregation concept, applicable to both feature selection and classification or regression problems.

Introduced by (Breiman, 2001), it utilizes both recursive partitioning (tree methodology) and bootstrap aggregating (bagging) methods. It combines multiple created decision trees, using several bootstrap samples yielded from the learning sample, choosing randomly at each node of a subset of predictor features.

Random forest is a powerful technique due to its simple implementation in classification and regression problems, able to achieve high prediction accuracy, with the capability of handling missing values, correlation and high dimensional data (Hapfelmeier and Ulm, 2013). In addition, random forest is more robust to noisy data and does not over fit as more tree is added, instead producing a limiting value of the generalization error (Friedman, 2001). Thus, it can be applied to a wide range of classification and regression problems even though it is non-linear and involves complex high-order interaction effects (Strobl *et al*., 2007).

As previously mentioned, the two steps in random forest consist of recursive partitioning and bagging. In step one, Classification And Regression Tree (CART) model building is used for recursive partitioning to construct the trees. The process involves growing and pruning. In CART, the trees are created through a growing process in which the sample data are split continuously in binary to generate subsets that are as homogeneous as possible. Creation continues until stopping criteria, such as maximum number of observations in each subset, are reached. In other cases, if the subset is homogenous, the growing process is stopped.

The splits are determined based on specified criteria which depend on the response type. The commonly applied binary and continuous responses in CART are the Gini index and residual sum of squares techniques respectively (Hapfelmeier and Ulm, 2013). CART differs from random forest in that during tree creation, a given number of input features are randomly chosen in each node and the best split is calculated only within this subset. The pruning process then creates sparser models that enhance predictive accuracy. Through this process, the created trees can be pruned back by evaluating the tree performance at different growth stages, using a cross-validation approach. However, in random forest, the pruning process is eliminated to ensure that all created trees are maximal trees.

Next, random forest applies the bagging method to further improve the tree methodology. During bagging, the selected trees are fitted to bootstrapped or subsampled data. The prediction outputs are calculated by averaging the values based on majority votes on each trees response which later enhances the prediction accuracy compared to single tree evaluation. According to (Breiman, 2001), bagging is applied to improve the accuracy of random features and, when used, provides continuous generalization error estimates for the combined ensemble of trees and is able to estimate the strength and correlation between features.

The combination of tree methodology and bagging in random forest provides a more diverse set of features for joint prediction, as the splits occur in random feature selection (Hapfelmeier and Ulm, 2013). Observations separate from the subsample used to create the respective tree are utilized to evaluate the accuracy of prediction. This provides a more realistic prediction performance when testing new data. These observations are called out-of-bag observations. In simple words, the subsample used to create trees is training data while out-of-bag observation is test data.

Random forest achieves higher prediction capabilities relative to other existing techniques, such as linear regression, artificial neural network, support vector machine and others. As explained in (Breiman, 2001), the growth of trees in this model relies on a random vector Θ where the tree predictor h (x, Θ) takes on numerical values corresponding to class labels. The values of the output result are numerical and it is assumed that the training data is drawn independently from the distribution of random vectors X and Y. The mean-squared generalization error $_{X, Y}$ for numerical predictor $h(x)$ is defined as:

$$E_{X,Y}\left(Y-h\left(X\right)\right)^2 \tag{9}$$

The random forest predictor is created by taking an average av over k of the trees $\{h\ (x, \Theta k)\}$. It is assumed that as more trees are grown, the mean-squared generalization error will converge and be defined as:

$$E_{X,Y}\left(Y-av_k h\left(X,\Theta\right)\right)^2 \to E_{X,Y}\left(Y-E_\Theta h\left(X,\Theta\right)\right)^2 \tag{10}$$

Thus, an average generalization of a tree can be summarized as follows:

$$PE\left(tree\right)=E_\Theta E_{X,Y}\left(Y-h\left(X,\Theta\right)\right)^2 \tag{11}$$

Then, an overall generalization error for all trees $PE(forest)$ is defined as:

$$PE\left(\ forest\right)=\rho\left(E_\Theta sd\left(\Theta\right)\right)^2 \tag{12}$$

where, ρ is a weighted correlation between the residual of Y-h (X,Θ) and Y-h (X,Θ') and sd a standard deviation of tree generalization error. Calculation of sd was defined in the following equation:

$$sd\left(\Theta\right)=\sqrt{E_{X,Y}\left(Y-h\left(X,\Theta\right)\right)^2} \tag{13}$$

Gradient Tree Boosting (GTB)

Gradient Tree Boosting (GTB), introduced, by (Friedman, 2001), develops a prediction model based on boosting and decision tree learning techniques. It is inspired by another statistical framework, called Adaptive Reweighting and Combining (ARC) algorithm, introduced by (Breiman, 1997). GTB is a stage-wise, additive framework that adopts numerical optimization methods to minimize the loss function of the predictive model which later enhances its predictive capabilities. The advantage of GTB lies in its ability to produce highly competitive, robust and interpretable solutions for both regression and classification problems (Friedman, 2001). In addition, an application of the boosting technique in GTB allows avoidance of overfitting problems when new independent data is added (Friedman, 2001).

Most decision tree methods tend to grow a single, large decision tree from available data, which causes overfitting and high variance. GTB's 'boosting' technique avoids such problems and minimizes variance in decision trees methods.

GTB preserves and sequentially grows the long learner tree, which iteratively learns and fixes the errors of previous iterations. Thus, the output result produced by GTB has low variance and error.

According to (Friedman, 2001) and (Jiang *et al*., 2007), from N training set $\{y_i, x_i\} i = 1N$ of known values (y, x), GTB's main objective is to find an estimation of function $F(x)$ that maps all x to y values where the loss function value of $L(y, F(x))$ is minimized for each iteration m. In the first step of GTB, the loss function $L(y, F(x))$ is defined first. Then, the initial value of $F0(x)$ is defined and its definition is shown in following equation:

$$F_0(x) = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, \rho) \tag{14}$$

$F_0(x)$ is an initial guess of successive increments ("steps" or "boosts") based on the sequence of the preceding steps of $F_m(x)$. ρ is the initial multiplier given by the line search of $F_m(x)$. For each successive $F_m(x)$, gradient descent boosting technique, using least square function as loss function for next $F_{m+1}(x)$, is applied and defined as:

$$Y_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, N \tag{15}$$

The output result produces a residual called 'pseudoresponses' γ_i that later is fitted with applied base or weak learner a_m. In GTB, the decision tree is applied as base or weak learner and defined as follows:

$$a_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \left[\gamma_i - \beta h(x_i; a)\right]^2 \tag{16}$$

In this step, β is a greedy stage-wise function that estimates $F(x)$ under the constraint that the step "direction" of $h(x_i; a)$ is a member of the parameterized class of functions $h(x; a)$. Next, multiplier ρ_m is computed, given by the line search for each respective $F_m(x)$ and shown in Equation 17:

$$\rho_m = \arg\min_{a} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i; a_m)) \tag{17}$$

Finally, the $F_m(x)$ estimation is updated as an output approximation defined by the following equation:

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m) \tag{18}$$

Each approximation output of $F_m(x)$, is then stored in a set of $F_M(x)$. The trained GTB are then tested, using the new test sample data to observe predictive performance.

In Fig. 1, GTB is a powerful artificial intelligence technique that is highly customizable and robust to the particular needs of application. Based on (Chandrasekar *et al*., 2015) and (Jiang *et al*., 2007), the advantages of GTB are:

- It produces highly competitive, robust and interpretable solutions for both regression and classification problems
- It avoids the overfitting problem when new independent data is added
- The output result produced has low variance and error
- It is flexible and diverse to data structure, performing well even with imbalanced data

Dragonfly Algorithm (DA)

Dragonfly Algorithm (DA) is a recently introduced, nature inspired metaheuristic optimization algorithm created by (Mirjalili, 2016), inspired by static and dynamic swarming behavior of dragonflies. DA adopts swarm intelligence concepts that mimic the unique social interaction of dragonflies in navigating, migrating, food searching and avoiding enemies. DA advantages include improvement of the initial random population, convergence towards the global optimum and the production of reliable results. DA is flexible as it is applicable in solving single-objective, multi-objective and discrete problems (Roberto *et al*., 2019; Mirjalili, 2016; Rahman and Rashid, 2019).

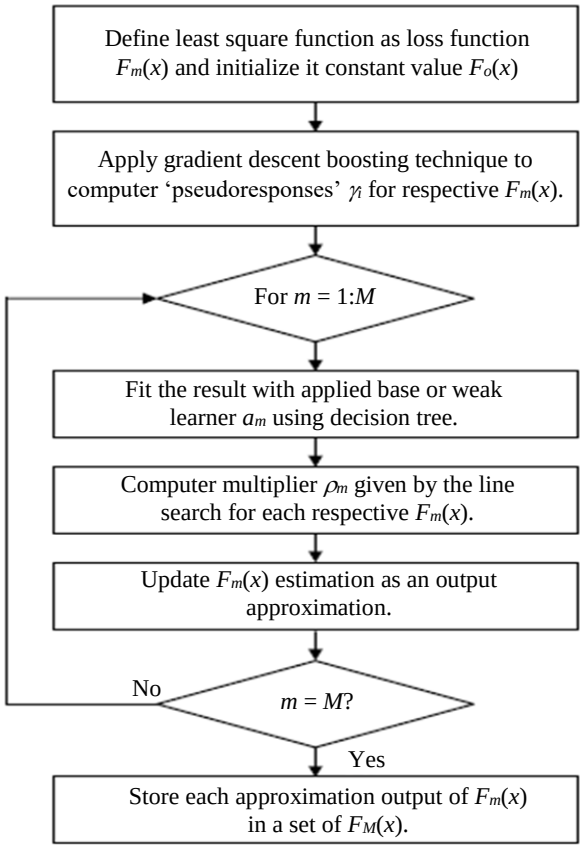

Fig. 1: Illustrates the GTB framework

As previously mentioned, dragonfly swarming behavior is primarily based on two features: Static and dynamic. These two swarming behaviors are similar to the two main phases, exploration (static) and exploitation (dynamic), in metaheuristic optimization concept. In static swarming, an individual or small group of dragonflies fly within a small area to search for food. The local movements and abrupt changes in the flying path of dragonflies match the characteristics of dragonfly's static swarming. The behavior of dragonflies in creating sub-swarms and moving to different places during static swarming similarly mirrors the exploration concept in DA. In dynamic swarming, dragonflies move in a massive swarm in one direction, migrating towards targeted locations. This dragonfly behavior is embraced by the exploitation concept in DA. Figure 2 illustrates the static and dynamic swarming behavior in DA.

According to (Mirjalili, 2016), like other nature inspired swarm intelligence based metaheuristic optimization algorithms, DA also adopts three basic principles: separation to avoid static collision between individuals in the neighborhood, alignment for velocity matching between individuals in the neighborhood and cohesion, which refers to the tendency of individuals to gravitate towards the center of mass of the neighborhood. The main survival tactics of dragonflies are the attempt to find food and avoidance of nearby enemies. In DA, the updating position of individual dragonfly behavior is based on five corrective patterns, including separation, alignment, cohesion, attraction to food and evasion of enemies. Each corrective pattern is mathematically modeled and calculated. Separation corrective patterns are calculated as follows:

$$S_i = -\sum_{j=1}^{N} X - X_j \tag{19}$$

where, N is the number of neighboring individuals, X is the current position of the individual and X_j is the position *j-th* of the neighboring individual. The alignment corrective pattern is calculated as follows:

$$A_i = \frac{\sum_{j=1}^{N} V_j}{N} \tag{20}$$

where, V_j is the velocity of the *j-th* neighboring individual. As for the cohesion corrective pattern, the calculation is:

$$C_i = \frac{\sum_{j=1}^{N} X_j}{N} - X \tag{21}$$

Attraction to food is then formulated as follows:

$$Fi = X^+ - X \tag{22}$$

where, X^+ is the position of the food source. The final corrective pattern, evasion of enemies, is defined as follows:

$$Ei = X^- + X \tag{23}$$

where, X^- is the position of the nearby enemies.

The combination of these five corrective patterns describes the overall behavior of each individual dragonfly in DA. During optimization, updating the position in DA requires two vectors, step ΔX and current position X. Step vector provides the movement direction of the dragonflies and it is defined as:

$$\Delta Xt + 1 = \left(sSi + \alpha Ai + cCi + fFi + eEi\right) + w\Delta Xt \tag{24}$$

where, s is separation weight, Si is separation of the *i-th* individual, a is alignment weight, Ai is alignment of the *i-th* individual, c is cohesion weight, Ci is cohesion of the *i-th* individual, f is food factor, Fi is food source of the i-th individual, e is enemy factor, Ei is position of enemy of the *i-th* individual, w is inertia weight and t is iteration number. After the step vector calculation is complete, the position vector of current iteration t is then calculated as follows:

$$Xt+1 = Xt + \Delta Xt + 1 \quad (25)$$

The position update in DA during optimization is calculated in an assumed neighborhood radius (boundary) for each individual dragonfly. Each of the position updates is performed within this neighborhood radius. Figure 3 illustrates the swarming behavior of each individual dragonfly in its respective neighborhood radius for finding food and avoiding enemies until the stopping criteria are achieved.

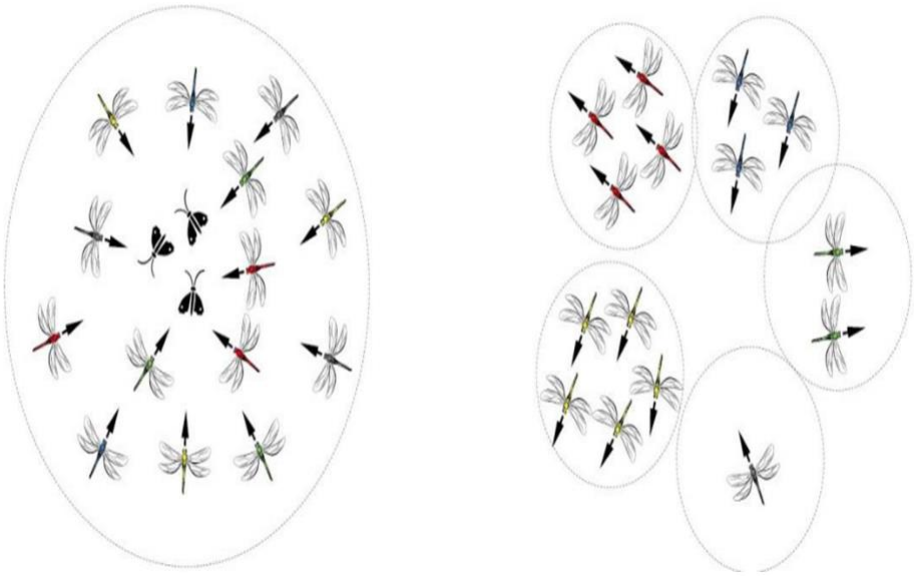

Fig. 2: Static and dynamic swarming behavior in DA (Mirjalili, 2016)

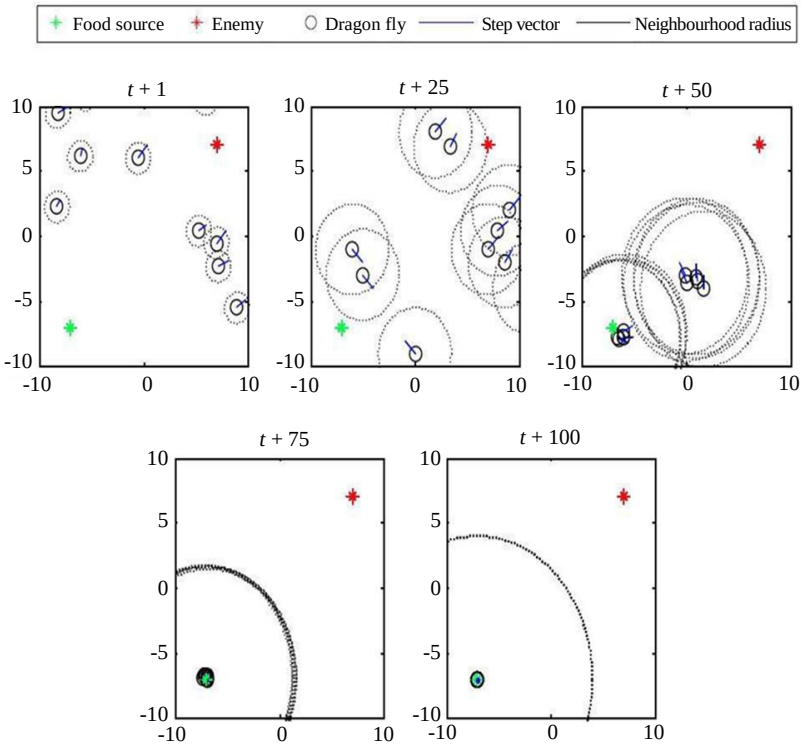

Fig. 3: Swarming behavior of each individual dragonfly in its respective neighborhood radius (Mirjalili, 2016)

In the case of no neighboring solutions being found, random walk (Levy flight) is employed for each individual dragonfly to fly around the search space (Mirjalili, 2016) This is done to improve randomness, stochastic behavior and exploration in DA. In this case, the update position of the dragonfly for the current iteration is defined as follows:

$$Xt+1 = Xt + Levy(d) \times Xt \tag{26}$$

where, d is dimension of the position vectors. The formula in calculating *Levy* is defined as:

$$Levy(d) = 0.01 \times \frac{r_1 \times \sigma}{\left|r_2\right|^{\frac{1}{\beta}}} \tag{27}$$

where, β is a constant value, r_1 and r_2 are two random number between 0 and 1. Parameter σ, is calculated as:

$$\sigma = \left(\frac{\Gamma(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}} \right)^{\frac{1}{\beta}} \tag{28}$$

where, Γ (x) = (x-1). Overall, the DA algorithm begins the process of optimization by constructing a set of random solutions based on a given problem. Initially, the position and step vector of the dragonflies' values are randomly defined within the lower and upper bounds of variables. In each iteration, both vectors are updated based on the five corrective patterns mentioned earlier. The update process is accomplished within the neighborhood radius by calculating the Euclidean distance between all dragonflies and a selected individual dragonfly. This process is continued iteratively until the stopping criteria are achieved.

Hybrid Techniques

In recent literature, researchers have extensively studied the integration of several artificial intelligence techniques in an attempt to improve the accuracy of existing forecasting models Such an approach is called a hybrid model, it combines different linear forecasting models, non-linear models, or both linear and non-linear models (Wu and Lu, 2012). The main objectives of the many types of hybrid models, introduced by different researchers, are to outperform other singular or independent forecasting models, overcome the limitations of current forecasting models and enhance forecast result accuracy. According to (Wu and Lu, 2012), artificial intelligence techniques are stochastic, not guaranteeing an optimal forecast result. Thus, hybridizing the forecasting model minimizes and overcomes the uncertainty problems which arise in artificial intelligence techniques.

One way of developing a hybrid model is to integrate other techniques in order to optimize the parameter input of the forecasting model. Input parameters in artificial intelligence techniques are considered critical because they heavily affect output results. Hence, it is compulsory to ensure that the provided parameters are appropriate and reliable. Several approaches have been introduced to overcome such problems in order to improve existing forecasting model accuracy, such as ANN (Hipp and Yates, 2011; Rather *et al*., 2017; Alwee *et al*., 2013; Yeh, 2013) and SVR (Wu *et al*., 2009; De Oliveira and Ludermir, 2014; Wu and Lu, 2012).

External Factors and Their Selection and Optimizer Algorithm

External Factors

Most criminology experts and researchers' study and observe the effects of several factors on criminal activities. This provides relevant insight into possible crime trends based on recent issues. In multivariate crime analysis, extensive studies have been conducted to observe the relationship between external factors and their impact on crime. Studying the influence of these factors in crime analysis offers significant benefit because, rather than a heavy dependence on past crime trends, crime occurrence patterns are affected by various factors such as social mistreatment, population densities and economic disadvantages. Crime analysis organizes external factors into three categories: Social, demographic and economic. Social factors include the impact of poverty, gender and race inequality on crime rate. Meanwhile, the study of demographic factors focuses on the impact of population (age, gender, race and immigration rate) and geographical (climate, urban area and crime hotspot) influence on crime trends. Economic factors subsume the relationship between crime trends and recent economic issues, such as unemployment, inflation and Gross Domestic Product (GDP) rate.

In the study of social factors, several researchers have observed a number of social influences on crime trends. Jain and Kumar (2007; Ridzuan Khairuddin *et al*., 2019; Triana and Retnowardhani, 2019) studied the relationship between maltreatment, family environment and social risk factors and the commission of crimes by children and adolescents. The study points out the possibilities of children and adolescents committing crimes based on maltreatment and family environment according to their age, race and gender. The study states that almost 3% of maltreated children will possibly commit a crime within an average of 6 years. Wu and Lu

(2012) studied the effect of immigration on crime in San Diego, USA. The study explores the impacts of immigration on neighborhood-level homicide trends using race disaggregated homicide victim data and community structural factors collected in three decennial census periods. The study results indicate that social disorganization in heavily immigrant cities possibly plays a major role in economic deprivation, which affects crime trends. Wu and Lu (2012) looked at the relationship between poverty rates and crime rates. The study observed that poverty rates, according to ethnic heterogeneity, affects various crime types. Based on the results discussed, it stated that murder crime is strongly affected by poverty.

As for demographic factors, researchers have studied the potential impacts of several demographic properties on crime patterns. Cook and Durrance (2013) investigated the relative role of demographic and substance use characteristics in nightlife violent crime involvement among Norwegian nightlife patrons. The study also incorporated individual measures of concurrent blood alcohol content level, population age and gender. The results state that patrons involved in violent crime were more likely to present with a blood alcohol level above 1% than those who had not been involved. Cook and Durrance (2013) studied the implication of shale rich regions (oil and gas discovery regions) with regards to regional crime rates in USA counties. The study also used population size, employment rate, Gross Domestic Product (GDP) and mining employment data to discover demographic effects on crime rate patterns. The results implied that there are significant changes in crime rates due to demographic shifts (population rising) in shale rich regions. Cook and Durrance (2013) estimated the impact of climate change on criminal activity in the USA. The study employed several demographic factors, including weather conditions and population densities and ages to observe their relationship with crime rate changes. The result proves that climate changes strongly influence criminal behavior, correlating with an increase in crime rate counts.

Among the aforementioned factors, the impact of economic conditions on crime trends and patterns have been most widely and extensively studied (Alwee *et al*., 2013). Cook and Durrance (2013) observed the impact of economic disadvantage on homicide. The study examined economic discrimination across age specific transitional periods from adolescence to adulthood. It also analysed the impact of economic disadvantages (unemployment, poverty and divorce rate) with other factors, including residential instability, family problems and population heterogeneity. Results reveal an increased likelihood of lethal violent crime among adolescents and young adults. Cook and Durrance (2013) looked at the effects of federal tax increases on crime rates. The study used alcohol tax rates and per capita alcohol consumption to point out their impacts on crime rates in the USA. The study results showed an incremental pattern over time, indicating that an increase in the price of alcohol, due to tax rate increases, heavily impacted crime rates.

Northrup and Klaer (2014) studied the effect of gross domestic product on violent crime. The study looked at the impact of GDP per capita, along with other factors such as graduation, unemployment and poverty rates on crime rate changes. The findings show that GDP does affect crime rates, with crime rate patterns changing over time. Cebula (2011) examined the effect of unemployment rates on property crime. The study also looked at several influencing sub-factors, including inflation rate, consumer price index, personal income and average earnings per job in evaluating crime rates changes. The result show that unemployment does affect crime rate pattern. In addition, in crime trend analysis, assessment is strengthened by the inclusion of other relevant sub-factors.

The ultimate objective of artificial intelligence technique is the automatic construction of an efficient model from the data it learns without the requirement of tedious and time-consuming human interference. As stated by (Ganjisaffar *et al*., 2011), the main difficulty in achieving such a goal is the requirement of proper parameter configuration, which allows learning algorithms to adapt to the particulars of a training set that fits application needs. Optimizing the parameters in artificial intelligence technique is not an easy task because improper parameter configuration leads to overfitting or under fitting problems that later affect the performance of corresponding artificial intelligence techniques. Therefore, the artificial intelligence technique, rather than attempting to predict the functional dependence between input and response variables, instead predicts the training data itself (Natekin and Knoll, 2013).

This issue is not new; several studies have been conducted over the past decade, looking to provide efficient solutions. The optimization of parameters through the selection of optimal values is very challenging because this optimization is conducted on not only one parameter but several other related parameters which control prediction performances. Thus, finding optimal combination values of the related parameters is challenging. Based on the literature study conducted, different researchers introduced different solutions to identify optimal parameter values that fit their application needs.

The Selection Algorithms

In multivariate time series models, the use of all available variables (factors or features) to develop the crime model is inefficient. According to (Han and Wang, 2009), although multivariate time series are able to discover more information about complex systems,

which can enhance the accuracy of forecasting, the use of too great a number of input variables leads to overfitting and poor generalization abilities. Therefore, the use of select variables in multivariate time series models is necessary to ensure that only important variables are chosen, thus avoiding the aforementioned problems. Feature selection (variable selection) is an effective solution when handling multivariate time series models because it is able to extract main features in the time series and at the same time minimizes the model inputs dimension (Han and Wang, 2009). Feature selection is very important to improve forecasting model accuracy and ensure it simplicity.

In forecasting using multivariate time series, feature selection is used to elucidate the strength of the relationship between dependent and independent variables. The dependent variables are target time series, which need to be forecast while the independent variables are external factors used to discover the new pattern of the time series data. Statistical correlational analysis is the most applied method in feature selection (Alwee *et al*., 2013). This method describes the degree of relationship between two variables and observes their correlation. It is limited in that it requires complete (sufficient) data to determine the significant relationships between variables, only considers linear relationships between variables and requires an assumption that the data are linear. In addition, a strong correlation does not imply a cause and effect relationship between variables. The real challenge of feature selection is that in real-world data, especially crime data, it is insufficient and non-linear in nature. Thus, applying a statistical method directly to real-world data is impractical. Hence, a suitable feature selection method is needed for handling crime data.

Existing Feature Selection Techniques

This literature study has identified already existing feature selection techniques for the identification of significant relationships between two variables. There are two types of feature selection techniques: The null hypothesis test and feature impotence method. Null hypothesis tests are formal statistical methods to observe statistical significance between two features based on a defined hypothesis statement. The selected features are tested to measure their relationship strength against this hypothesis statement. In the feature importance method, the relationship strength between selected features is measured based on calculated importance values. The higher the importance value, the stronger the relationship between selected features. If the importance value is nearly 0 or negative, it indicates that there is no significant relationship between selected features. Two examples of feature selection techniques using a null hypothesis test are the F-Test and Student T-Test. The F-Test is a statistical hypothesis test used to find statistical significance between two features means by observing whether the variances of two features are equal or not. It was introduced by Ronald A. Fisher in 1920 and remains useful in determining the best variables that fit to the target sample. The F-Test is a popular feature selection technique due to its simplicity and ability to handle regression problems. It calculates the critical value to determine the significance between two variables and is tested under the null hypothesis test. The Student T-Test is a statistical hypothesis test used to find statistical significance between two variables means based on t-distribution under the null hypothesis. It was introduced by William Sealy Gosset in 1908 and has become one of the most popular feature selection techniques due to its simplicity and robustness. It computes the T-score between predictor features and response features to determine whether they are statistically significant or not.

Meanwhile, two examples of feature selection techniques using the feature importance method are ReliefF and Neighborhood Component Analysis (NCA). ReliefF is a feature selection technique introduced by (Robnik-Šikonja and Kononenko, 2003) to handle noise and multiclass datasets and it is an extension of the relief algorithm proposed by Kira and Rendell in 1992 (Han and Wang, 2009). ReliefF is able to estimate feature quality correctly by observing the dependencies between features in various conditions. It is popular because it is able to handle classification and regression problems, as well as low bias and is robust to feature interaction and diverse to data structure whether incomplete, binary, or continuous. It is useful in the preprocessing data phase to select relevant features before the model is trained. NCA is a non-parametric feature selection technique that applies embedded methods to select relevant features that are able to improve prediction and classification accuracy. Its objective is to find a linear transformation from a set of features by maximizing a stochastic variant of the leave-one-out K-Nearest Neighbor (KNN) score.

The application of feature selection techniques in improving forecasting performance in crime forecasting models is still limited and rarely applied. Thus, in this study, the appropriate application of feature selection is proposed to identify and select the important external factors that significantly affect crime. External factors are the features or variables that will be assessed to observe its relationship with corresponding crime types. By identifying and selecting significant factors and eliminating irrelevant factors, proposed crime forecasting model performances can be improved. In this study, Neighborhood Component Analysis (NCA) is considered in selecting the significant external factors that influence crime.

Neighborhood Component Analysis (NCA)

Neighborhood Component Analysis (NCA) is a non-parametric feature selection technique that applies

embedded methods to select relevant features that improve prediction and classification accuracy. It was introduced by (Yang *et al*., 2012) with the motivation of improving the KNN algorithm. NCA is a feature weighting method based on the nearest neighbor approach that applies gradient ascent technique to maximize the expected leave-one-out accuracy with a regularization term (Yang *et al*., 2012). The main objective of NCA is to discover a weighting vector w, which it is used to determine the relevant feature by optimizing the nearest neighbor to solve classification or regression problems. The advantages of NCA lie in its ability to minimize overfitting during data training and its insensitivity to features number (Yang *et al*., 2012).

First, the samples of training set, $T = \{(x_1,y_1),\ldots,(x_i,y_i),\ldots,(x_N,y_N)\}$ are defined where *xi* is a *d*-dimensional of the predictor feature vector, y_i is a target feature vector and *N* is the number of samples. Next, the weighting vector *w* is defined based on weighted distance between two samples, x_i and x_j:

$$D_w\left(x_i,x_j\right)=\sum_{l=1}^{d} w_l^2\left|x_{il}-x_{jl}\right| \tag{29}$$

where, w_l is a weight related to *lth* feature. Then, leave-one-out classification accuracy on training set *T* is maximized so that the applied nearest neighbor technique can perform well. Next, a probability distribution is applied to approximate the reference point that determines the neighbor. Such an approach is applied because the true leave-one-out accuracy that selects neighbor as the reference point is a non-differentiable function Yang *et al*., 2012). The applied probability distribution of x_i that selects x_j as reference point is defined as follows:

$$P_{ij}=\begin{cases}\dfrac{k\left(D_w\left(x_i,x_j\right)\right)}{\sum_{x\neq i}k\left(D_w\left(x_i,x_x\right)\right)}, if\ i\neq j\\ 0, if\ i=j\end{cases} \tag{30}$$

where, $k(z) = \exp(-z/\sigma)$ is a kernel function and σ is a kernel width that affects each selected reference point probability. The selection of reference point based on kernel width *o* is based on two definitions The first definition is states that if $\sigma\rightarrow 0$, only the nearest neighbor of the query sample can be selected. The second definition states that If $\sigma\rightarrow+\infty$, all points have a fair chance to be selected as the reference point instead of query point. Based on these definitions, the query point *xi* probability is correctly defined as:

$$P_i=\sum_j y_{ij}P_{ij} \tag{31}$$

where, $y_{ij} = 1$ if and only if $y_i = y_j$ and $y_{ij} = 0$ otherwise. Based on these approximations, leave-one-out accuracy can be written as:

$$\xi(w)=\frac{1}{N}\sum_i P_i=\frac{1}{N}\sum_i\sum_j y_{ij}P_{ij} \tag{32}$$

Equation 32 is further improved by applying a regularization approach, to alleviate overfitting in feature selection based on the following defined objective function:

$$\xi(W)=\sum_i\sum_j y_{ij}P_{ij}-\lambda\sum_{l=1}^{d} w_l^2 \tag{33}$$

Note that the coefficients 1/N in Equation 32 are removed since the final solution is the same due to changes made by the regularization parameter. The value can be configured to improve NCA's performance to detect relevant features where > 0 (Yang *et al*., 2012). Finally, the derivation of $\xi(w)$ with respect to *wl* is computed as follows:

$$\begin{aligned}\frac{\partial\xi(w)}{\partial w_l}&=\sum_i\sum_j y_{ij}\left[\frac{2}{\sigma}P_{ij}\left(\sum_{x\neq i}P_{ix}\left|x_{il}-x_{xl}\right|-\left|x_{il}-x_{jl}\right|\right)w_l\right]-2\lambda w_l\\&=\frac{2}{\sigma}\sum_i\left(P_i\sum_{x\neq i}P_{ix}\left|x_{il}-x_{xl}\right|-\sum_j y_{ij}P_{ij}\left|x_{il}-x_{jl}\right|\right)w_l-2\lambda w_l\\&=2\left(\frac{1}{\sigma}\sum_i\left(P_i\sum_{j\neq i}P_{ij}\left|x_{il}-x_{jl}\right|-\sum_j y_{ij}P_{ij}\left|x_{il}-x_{jl}\right|\right)-\lambda\right)w_l\end{aligned} \tag{34}$$

As previously mentioned, the regularization parameter is very important in alleviating overfitting problems in feature selection in NCA. The optimal regularization parameter value is able to reduce the generalization error in NCA and thus improve forecasting performances.

New Approach in Selecting External Factors

The main principle of GTB is the construction of a new base or weak learner, meant to be highly correlated with gradient of loss function, which is associated with the whole ensemble (boosting and decision tree). The function of applied loss function in GTB is to consecutively fit new models in order to provide more accurate prediction (Guelman, 2012). Hence, loss function plays critical roles that determine GTB predictive capabilities and performances. As mentioned before, GTB uses least square function as a loss function to consecutively minimize 'pseudoresponses' value (error-fitting) over the response variable. Generally, the distribution of response variables varies and is not constant (Natekin and Knoll, 2013). Given that loss function is reflected on it, it is recommended to consider other potential mathematical functions to be used as loss function in GTB instead of only depending on least square function.

The appropriate application of loss function is beneficial as it provides flexibility in model designs that fit to different application needs (Guelman, 2012). Hence, such an approach provides robustness to GTB that is suitable and fit for the proposed crime forecasting

model. A study of the literature shows that there are different types of applied mathematical functions that can be selected as loss functions for GTB and are suitable for specific application needs. Each type of loss function is suitable for a specific task, whether it is used to solve classification or regression problems that fit to the target model (Ridgeway, 2013).

Conclusion and Future Work

In crime analysis, most crime models are based on time series data. Univariate and multivariate analysis are the two methods used in analyzing this type of data. Univariate analysis involves only single time series data to develop the forecasting model, while multivariate analysis employs more than one set of time series data in model development. In comparison, multivariate analysis provides better forecasting accuracy relative to univariate analysis because of its ability to discover patterns not previously seen (Alwee *et al*., 2013; Northrup and Klaer, 2014).

Based on the conducted literature study, several types of crime forecasting models have been introduced, including the statistical model, artificial intelligence model and hybrid model, which is a mixture of both. Currently the artificial intelligence model is favoured by most researchers in forecasting crime. This is due to its adaptability in handling the non-linear nature of most real-world data, unlike the statistical model, which fails to do so. Within the artificial intelligence model, Gradient Tree Boosting (GTB) is advantageous because of its avoidance of overfitting problems when new independent data is added, flexibility to data structure as it performs well even with imbalanced data and the low variance and error produced in the output result (Nguyen *et al*., 2017).

In GTB, loss function plays critical roles that determine its predictive capabilities and performances. As previously mentioned, GTB uses least square function as a loss function to consecutively minimize its 'pseudoresponses' value (error-fitting) over the response variable. Generally, the distribution of response variables is varied and not constant (Natekin and Knoll, 2013). Hence, it is possible to implement other standard mathematical functions in replacing least square function in GTB, improving and enhancing GTB performance capabilities in forecasting crime.

Findings also indicate that crime occurrence is influenced by several external factors, including social, demographic and economic. The influences of these factors on crime occurrence have been extensively studied to determine the impact of different issues that implicitly or explicitly affect crimes. However, not all factors have a significant effect on crime occurrence. Hence, analysis is required to observe, identify and select the significant factors that directly influence crime occurrence. One approach is the implementation of feature selection techniques to discover significant relationships between observed factors and crime rates. Neighborhood Component Analysis (NCA) is an effective feature selection technique, used in identifying significant factors that strongly influence crime. This is because NCA is capable of learning feature weighting vector by maximizing the expected leave-one-out regression accuracy with a regularization term that improves forecasting accuracy (Yang *et al*., 2012).

Another issue is the sensitivity of both GTB and NCA to parameter input configuration (Zhang, 2003; Nguyen *et al*., 2017). The most commonly applied solution to this problem is the implementation of metaheuristic optimization algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC) in identifying optimal parameter values that fit to application needs. Among the introduced optimization algorithms, Dragonfly Algorithm (DA) is shown to be a promising solution because of its flexibility in solving most optimization problems and ability to improve the initial random population, converge towards the global optimum and produce reliable results (Mirjalili, 2016). Thus, the optimization of GTB and NCA parameter values improves their overall performance capabilities.

Overall, the purpose of this study is to propose an effective multivariate crime forecasting model to accurately forecast crime rate data. The proposed crime forecasting model implements Gradient Tree Boosting (GTB) to forecast crime rates based on crime type. It is also equipped with Neighborhood Component Analysis (NCA) to identify and select significant external factors that influence crime. The external factors that affect crime used in this study include unemployment, immigration, population, consumer price index, gross domestic product, consumer sentiment index, poverty, inflation and tax. GTB is further enhanced by implementing a new loss function to improve its forecasting performance. In addition, both NCA and GTB parameter values are optimized using DA, which significantly improves their performance capabilities. The parameters to be optimized in GTB include Number of Trees (NoT), Size of Individual Trees (SoIT) and learning rate (LR), while the parameter to be optimized in NCA is the regularization parameter λ.

In conclusion, the proposed model is expected to produce an accurate forecast result and outperform other existing crime forecasting models, such as Artificial Neural Network (ANN), Support Vector Regression (SVR) and Random Forest (RF). Hence, it serves as an effective tool in forecasting crime rates and provides important insight into potential future crime trends for authorities to plan effective crime prevention measures.

Based on the discussion of the dataset, techniques and factors involved in crime forecasting of violence and

property, it is suggested that future work should be focused on the American dataset, existing external factor, hybrid model of dragonfly and support vector regression with eGTB parameter optimizer. This suggestion is justified because the American dataset is reliable, DA and SVM have better performance and have never been used in crime forecasting as far as literature review is concerned. Besides the proposed techniques, an optimizer should be considered to give better performance and accurate results by selecting the optimal external factors used in the forecasting.

Acknowledgement

I would like to show my massive appreciation for all supervisors of my Ph.D. Without them the work would not be possible. Their contribution was the key for publishing this paper.

Author's Contributions

Rebaz Nabi: This work has been accomplished as one of the taken papers from the Ph.D. Dissertation. He is the core contributor of this work.

Soran Saeed: This author was core supervisor of the thesis and work extensively towards preparing the article.

Habibollah Harron: This author was second supervisor of the thesis and work extensively towards preparing the article.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved

References

Alwee, R., S. Mariyam, H. Shamsuddin and R. Sallehuddin, 2013. Hybrid support vector regression and autoregressive integrated moving average models improved by particle swarm optimization for property crime rates forecasting with economic indicators.

Amroune, M., T. Bouktir and I. Musirin, 2018. Power system voltage stability assessment using a hybrid approach combining dragonfly optimization algorithm and support vector regression. Arabian J. Sci. Eng., 43: 3023-36. DOI: 10.1007/S13369-017-3046-5

Arora, S., H. Singh, M. Sharma, S. Sharma and P. Anand, 2019. A new hybrid algorithm based on grey wolf optimization and crow search algorithm for unconstrained function optimization and feature selection. IEEE Access, 7: 26343-264361. DOI: 10.1109/ACCESS.2019.2897325

Baliyan, A., K. Gaurav and S.K. Mishra, 2015. A review of short term load forecasting using artificial neural network models. Proc. Comput. Sci., 48: 121-25. DOI: 10.1016/J.PROCS.2015.04.160

Breiman, L., 1997. Arcing the edge. Statistics, 4: 1-14.

Breiman, L., 2001. Scalable parcel-based crop identification scheme using sentinel-2 data time-series for the monitoring of the common agricultural policy. Remote Sens., 10: 5-32.

Bye, E.K., 2007. Alcohol and violence: Use of possible confounders in a time-series analysis. Addiction, 102: 369-76. DOI: 10.1111/j.1360-0443.2006.01701.x

Cebula, R.J., 2011. Revisiting property crime and economic conditions: An exploratory study to identify predictive indicators beyond unemployment rates-comment. Soc. Sci. J., 49: 314-16. DOI: 10.1016/J.SOSCIJ.2010.07.015

Cesario, E., C. Catlett and D. Talia, 2016. Forecasting crimes using autoregressive models. Proceedings of the 14th International Conference on Dependable, Autonomic and Secure Computing, 14th International Conference on Pervasive Intelligence and Computing, 2nd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress, Aug. 8-12, IEEE Xplore Press, Auckland, New Zealand, pp: 795-802. DOI: 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.138

Chandrasekar, A., P. Kumar and A.S. Raj, 2015. Crime prediction and classification in San Francisco city we are dealing with the problem of crime classification in san.

Chen, P., H. Yuan and X. Shu, 2008. Forecasting crime using the ARIMA model. Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, Oct. 18-20, IEEE Xplore Press, Shandong, China, pp: 627-30. DOI: 10.1109/FSKD.2008.222

Chen, S.M. and K. Tanuwijaya, 2011. Multivariate fuzzy forecasting based on fuzzy time series and automatic clustering techniques. Expert Syst. Applic., 38: 10594-10605. DOI: 10.1016/J.ESWA.2011.02.098

Cheng, C.H., G.W. Cheng and J.W. Wang, 2008. Multi-attribute fuzzy time series method based on fuzzy clustering. Expert Syst. Applic., 34: 1235-1242. DOI: 10.1016/j.eswa.2006.12.013

Cook, P.J. and C.P. Durrance, 2013. The virtuous tax: lifesaving and crime-prevention effects of the 1991 federal alcohol-tax increase. J. Health Econom., 32: 261-267. DOI: 10.1016/j.jhealeco.2012.11.003

De Oliveira, J.F.L. and T.B. Ludermir, 2014. A distributed PSO-ARIMA-SVR hybrid system for time series forecasting. Proceedings of the International Conference on Systems, Man and Cybernetics, Oct. 5-8, IEEE Xplore Press, San Diego, CA, USA, pp: 3867-3872. DOI: 10.1109/SMC.2014.6974534

Foucquier, A., S. Robert, F. Suard, L. Stéphan and A. Jay, 2013. State of the art in building modelling and energy performances prediction: A review. Renewable Sustainable Energy Rev., 23: 272-288. DOI: 10.1016/j.rser.2013.03.004

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Statist., 29: 1189-1232. DOI: 10.1214/aos/1013203451

Ganjisaffar, Y., R. Caruana and C.V. Lopes, 2011. Bagging gradient-boosted trees for high precision, low variance ranking models. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, (DIR' 11), pp: 85-94. DOI: 10.1145/2009916.2009932

Guelman, L, 2012. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst. Applic., 39: 3659-67. DOI: 10.1016/J.ESWA.2011.09.058

Han, M. and Y. Wang, 2009. Analysis and modeling of multivariate chaotic time series based on neural network. Expert Syst. Applic., 36: 1280-90. DOI: 10.1016/J.ESWA.2007.11.057

Hapfelmeier, A. and K. Ulm, 2013. A new variable selection approach using random forests. Comput. Stat. Data Anal., 60: 50-69. DOI: 10.1016/J.CSDA.2012.09.020

Hipp, J.R. and D.K. Yates, 2011. Ghettos, thresholds and crime: does concentrated poverty really have an accelerating increasing effect on crime. Criminology, 49: 955-990. DOI: 10.1111/j.1745-9125.2011.00249.x

Jain, A. and A.M. Kumar, 2007. Hybrid neural network models for hydrologic time series forecasting. Applied Soft Comput. J., 7: 585-92. DOI: 10.1016/j.asoc.2006.03.002

Jiang, H., K. Huang and R. Zhang, 2007. Field Support Vector Regression. In: Neural Information Processing, Liu D., S. Xie, Y. Li, D. Zhao and E.S. El-Alfy (Eds.), Springer, Cham, ISBN-13: 978-3-319-70086-1, pp: 699-708.

Khashei, M. and M. Bijari, 2011. A novel hybridization of artificial neural networks and arima models for time series forecasting. Applied Soft Comput. J., 11: 2664-2675. DOI: 10.1016/J.ASOC.2010.10.015

Mcclendon, L. and N. Meghanathan, 2015. Using machine learning algorithms to analyze crime data. Mach. Learn. Applic. Int. J., 2: 1-12. DOI: 10.5121/mlaij.2015.2101

Mirjalili, S., 2016. Dragonfly algorithm: A new meta-heuristic optimization technique for solving single-objective, discrete and multi-objective problems. Neural Comput. Applic., 27: 1053-1073. DOI: 10.1007/s00521-015-1920-1

Natekin, A. and A. Knoll, 2013. Gradient boosting machines, a tutorial. Frontiers Neurorobotics.

Nguyen, T.T., A. Hatua and A.H. Sung, 2017. Building a learning machine classifier with inadequate data for crime prediction. J. Adv. Inform. Technol., 8: 141-47.

Northrup, B. and J. Klaer, 2014. Effects of GDP on violent crime.

Rahman, C.M. and T.A. Rashid, 2019. Dragonfly algorithm and its applications in applied science survey. Comput. Intell. Neurosci.

Rather, A.M., V.N. Sastry and A. Agarwal, 2017. Stock market prediction and portfolio selection models: A survey. Opsearch, 54: 558-579. DOI: 10.1007/s12597-016-0289-y

Ridgeway, G., 2013. Generalized boosted models: A guide to the GBM package. Computer, 1: 1-12.

Ridzuan Khairuddin, A., R. Alwee and H. Haron, 2019. A review on applied statistical and artificial intelligence techniques in crime forecasting. IOP Conf. Series: Mater. Sci. Eng.

Roberto, B., P. Velloso and C.F. Dorneles, 2019. Web Engineering. Proceedings of the 19th International Conference on Web Engineering, Jun. 11-14, Springer, Daejeon, South Korea, pp: 3-18. DOI: 10.1007/978-3-030-19274-7

Robnik-Šikonja, M. and I. Kononenko, 2003. Robnik-šikonja-kononenko2003_Article_Theoretical and empiricalanalysi. Pdf. Mach. Learn., 53: 23-69. DOI: 10.1023/A:1025667309714

Song, J., J. Wang and H. Lu, 2018. A novel combined model based on advanced optimization algorithm for short-term wind speed forecasting. Applied Energy, 215: 643-658. DOI: 10.1016/J.APENERGY.2018.02.070

Strobl, C., A.L. Boulesteix, A. Zeileis and T. Hothorn, 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinform.

Triana, Y.S. and A. Retnowardhani, 2019. Enhance interval width of crime forecasting with arima model-fuzzy alpha cut. Telkomnika Telecommun. Comput. Electron. Control, 17: 1193-1201. DOI: 10.12928/telkomnika.v17i3.12233

Wu, C.H., G.H. Tzeng and R.H. Lin, 2009. A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. Expert Syst. Applic., 36: 4725-35. DOI: 10.1016/J.ESWA.2008.06.046

Wu, C.H., G.H. Tzeng, Y.J. Goo and W.C. Fang, 2007. A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. Expert Syst. Applic., 32: 397-408. DOI: 10.1016/j.eswa.2005.12.008

Wu, J. and Z. Lu, 2012. A novel hybrid genetic algorithm and simulated annealing for feature selection and kernel optimization in support vector regression. Proceedings of the 5th International Conference on Advanced Computational Intelligence, Oct. 18-20, IEEE Xplore Press, Nanjing, China, pp: 999-1003. DOI: 10.1109/ICACI.2012.6463321

Xiao, H., W. Pei, Z. Dong, L. Kong and D. Wang, 2018. Application and comparison of metaheuristic and new metamodel based global optimization methods to the optimal operation of active distribution networks. Energies.

Yang, W., K. Wang and W. Zuo, 2012. Neighborhood component feature selection for high-dimensional data.

Yeh, W.C., 2013. New parameter-free simplified swarm optimization for artificial neural network training and its application in the prediction of time series. IEEE Tran. Neural Networks Learn. Syst., 24: 661-65. DOI: 10.1109/TNNLS.2012.2232678

Zhang, P.G., 2003. Time series forecasting using a hybrid arima and neural network model. Neurocomputing, 50: 159-175. DOI: 10.1016/S0925-2312(01)00702-0