Journal
of Science
Research

Forecasting infectious disease outbreak using support vector regression (SVR) case study: measles (rubeola)

Akomolafe O.P.[1], Adewuyi P.[2] and Nzenwata U.J.[1]
[1]Department of Computer Science, University of Ibadan, Oyo State, Nigeria.
[2]Department of Medicine & Health Science, Federal, University of Agriculture, Abeokuta, Nigeria.
Corresponding author: akomspatrick@yahoo.com; peter.adewuyi@gmail.com; uchennajerry@gmail.com

Abstract

Disease outbreak forecasting, provides warning that a certain amount of disease may occur at a particular time in the future. This research work uses measles, which is a highly contagious disease caused by the measles virus *Morbillivirus* as a case study. It has been problematic detecting the outbreaks of measles, which leads to high childhood mortality rate with either little or no response from the public health workers. Therefore, there is the need to forecast measles outbreaks to assist the public health workers facilitate preventive measures in Oyo state. By training a machine learning algorithm, Support Vector Regression (SVR), using the past measles outbreak records (2008-2015), obtained from the ministry of health, Oyo state, Ibadan; thirty-three models representing the thirty-three local governments in Oyo state were generated. Three features were extracted which are, Moving Average (MA), Statistic, and Relative Strength Indicaor. The result of this research project returned a Boolean value which depends on the set outbreak threshold. Mean Squared Error (MSE) and Mean Relative Error (MRE) were the metrics used to measure the performance of the algorithm. Another parameter of significance is the window size, which represents the number of previous data selected in order to estimate each feature from the measles record data. Therefore, it can be concluded that the window size value affected the training time of the algorithm and the efficiency of the models generated. The results of this research can be used as a tool to facilitate the preparedness against Measles outbreak ahead of time.

Keyword. Moving Average (MA), Statistic, Support Vector Regression, Relative Strength Indicator, Windows size.

Introduction

Predictive analytics is an approach in which data are analyzed for meaningful patterns that can provide actionable insights, which can in turn be used to enhance preventive actions and provide cost effective management that improves the domain products (examples in health care include patient outcomes, patient satisfaction, resources allocation, system re-structuring, and others) [1].

Emerging infectious diseases pose a growing threat to human populations. Many of the world's epidemic diseases (particularly those transmitted by intermediate hosts) are known to be highly sensitive to long-term changes in climate and short-term fluctuations in the weather [2]. Existing mechanisms for infectious disease surveillance and response are inadequate to meet the increasing needs for prevention, detection, reporting and response [3]. The ability to forecast epidemics will provide a mechanism for governments and health-care services to respond to outbreaks in a timely fashion, enabling the impact to be minimized and limited resources to be saved.

Due to the emergence and re-emergence of infectious diseases with pandemic potentials, there has recently been much interest in their analysis [4]. Currently, a large amount of infectious disease data is routinely collected by laboratories, health care providers and government agencies in an effort to increase the understanding of their evolution and predict, detect, prevent and manage the outbreak of infectious diseases. With this in view, Measles disease is used as a case study in this literature.

In this paper, a regressive study of past measles outbreak records was established, and used for the forecast of measles outbreak on a monthly basis. The data gathered is a monthly record of the number of outbreaks in each local government of Oyo state; the time series data was transformed into features that represent the factors that could affect measles outbreak Therefore, we present a forecasting methodology best fit for the extracted features with regression models where we propose Support Vector Regression for model construction. Our methodology however, generates models for each local government.

The remainder of this study is organized as follows. Section II provides a literature review on related works, Time series, and support vector regression. In Section III we developed the proposed forecasting methodology. Section IV presents the results derived by our methodology as well as the discussion. Section V concludes this work with recommendations.

Literature review

Review of related works

Compared Various Forecasting Methods for Auto correlated Time Series. Two machine learning methods, Artificial Neural Network (ANN) and Support Vector Machine (SVM), and a traditional approach, the Autoregressive Integrated Moving Average (ARIMA) model, were utilized to predict the demand for consumer products. The training data used were the actual demand of six different products from a consumer product company in Thailand. Initially, each set of data was analyzed using Ljung-Box-Q statistics to test for autocorrelation. Afterwards, each method was applied to different sets of data. The results indicated that the SVM method had a better forecast quality (in terms of MAPE: Mean Absolute Percentage Error) than ANN and ARIMA in every category of products. The gap in this literature emphasized that the autocorrelation structure of the data used has effects on the performance of the SVM algorithm, though the SVM still appears to be the best amongst the three methods.

A High-Priority Infectious Disease Surveillance Regions forecast using a socioeconomic model to forecast national annual rates of infectious disease outbreaks. A multivariate mixed-effects Poisson model was constructed of the number of times a given country was the origin of an outbreak in a given year. The dataset included 389 outbreaks of international concern reported in the World Health Organization's Disease Outbreak News from 1996 to 2008. The initial full model included 9 socioeconomic variables related to education, poverty, population health, urbanization, health infrastructure, gender equality, communication, transportation, and democracy, and 1 composite index. Population, latitude, and elevation were included as potential confounders. The initial model was pared down to a final model by a backwards elimination procedure. The dependent and independent variables were lagged by 2 years to allow for forecasting future rates [6].

There was a large proportion of missing values in our socioeconomic data, as expected, as not all countries collect and/or report these data or do so every year.

The Mathematical model of the dynamics of measles in New Zealand was developed in 1996. The model successfully predicted an epidemic in 1997 and was instrumental in the decision to carry out an intensive MMR (measles-mumps rubella) immunization campaign in that year. While the epidemic began some months earlier than anticipated, it was rapidly brought under control, and its impact on the population was much reduced. In order to prevent the occurrence of further epidemics in New Zealand, an extended version of the model had since been developed and applied to the critical question of the optimal timing of MMR immunization [7].

A hybrid methodology that exploits the unique strength of the Seasonal Autoregressive Integrated Moving Average (SARIMA) model and the SVM model in forecasting seasonal time series. The seasonal time series data of Taiwan's machinery industry production values were used to examine the forecasting accuracy of the proposed hybrid model. The forecasting performance was compared among three models, i.e., the hybrid model, SARIMA models and the SVM models, respectively.

Among these methods, the Normalized Mean Square Error (NMSE) and the Mean Absolute Percentage Error (MAPE) of the hybrid model were the lowest. The hybrid model was also able to forecast certain significant turning points of the test time series. The experimental results showed that the hybrid model (SARIMASVM2) is superior to the individual models (SARIMA and SVM models) for the test cases of the production value of the Taiwanese machinery industry. The NMSE and MAPE were all lowest for the hybrid model. The hybrid model also outperformed other models in terms of overall proposed criteria, including NMSE, MAPE, and turning point forecasts. Overall, the results obtained by the hybrid models were superior to those obtained using the individual models, in terms of both prediction errors and directional change detectability [8].

This method is only effective using time series model and cannot be used where models are generated dynamically.

Reviewed the application of statistical models to the outbreaks of measles epidemic. The epidemiological characteristics was paramount to this research and assessed the extent to which those characteristics either aid or hinder modeling. The developed models were turned to simulate geographical spread. A distinction was drawn between process-based and time series models. They provided applications from work, by using Icelandic data. Finally they considered the forecasting potential of the models described.

A new approach was proposed to forecasting based on the Bayesian principles of information theory and called the Poisson - gamma single - state model. In the research, a two-state version of the Poisson - gamma model was formulated by considering the uncertainty not only in the parameters but also in the model itself. That model was particularly useful for modeling epidemic data such as measles by considering two different situations of the generating process at each time point [10].

In the study of predicting Dengue Hemorrhagic Fever (DHF) in Thailand, an automatic prediction system for DHF is proposed by utilizing entropy technique and ANN. Entropy is used to extract the relevant information that affects the prediction accuracy. Later, the supervised neural network is applied to predict future DHF outbreak. Result obtained revealed that, by applying entropy technique, it would yield a better result as the entropy technique produces 85.92% accuracy while only 78.16% when entropy technique is not applied [11].

Time series

Defined mathematically as a time dependent sequence

$$X_1, X_2, X_3, \dots, X_N \qquad (1)$$

Where *1, 2, 3… N* depicts time steps and are assumed to be equally spaced [12].

Time Series can be classified into: deterministic time series and stochastic time series. - Time series which can be expressed as a known function, such as $X_t = f(t)$ is said to be deterministic time series.
The sequence of random variables
$\{X_t : \mathrm{t} = 0, \pm1, \pm2, \pm3, \dots\}$ is called a stochastic process and serves as a model for an observed time series. Time series is said to be stochastic time series if it can be expressed as $X_t = (X_t)$, where X is a random variable. Here the mean function is defined by $\mu_t = E(X_t)$ for $t = 0, \pm1, \pm2, \pm3, \dots$ [12].

Support vector regression

Support Vector Machine (SVM)[13] algorithm is also known as large margin classifier because it tries to find the best separation margin for the data. It does this by looking for the margin with the highest size within the two data points being. Although SVM is mostly used for classification problem there is a modification of its type used for regression problem called Support Vector Regression (SVR).

Given a training sample (also called training data) $(x_1, y_1), \dots.., (x_n, y_n)$ where x_i is an element of X, the input value, and yi is an element Y,
the output value such that i ϵ {1,…,n}
where n is the number of training examples.

The basic idea of SVR is to find a function:

$$f(x) = w.x + b \qquad (2)$$

with at most Ɛ-deviation from the target value y. Where x, w ϵ R^m where m is the number of features, that is the number of column if the sample data is represented inform of a table and w is the coefficient of x. This means that x and w are vectors while the statement above can be written mathematically as:

$$f(x_i) - y \leq \mathcal{E} \qquad (3)$$

where Ɛ represent a very small value.
Also

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + b \quad (4)$$

the objective of the algorithm is to find the values w and b such that the condition in eqn (1) above is satisfied.

To satisfy the condition, the algorithm have to minimize the value $\frac{1}{2}||w||^2$. Sometimes, it might be impossible to find a function f(x) that actually satisfies the condition in equation (2). In order to cater for this condition, there is a need to allow for some error by introducing a slack variable $£_i$, $£_i^*$; and as a result of this slack variable the minimization equation becomes:

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(£i + £i\,*) \qquad (5)$$

With respect to

$$f(x_i) - y_i \leq \mathcal{E} + £_i^* \text{ and}$$
$$y_i - f(x_i) \leq \mathcal{E} + £_i$$

where $£_i, £_i^* > 0$.

The minimization in this value is done using a quadratic programming to construct lagrange function.

After the parameter has been derived using the training data, further prediction is performed using the function in eqn (3) with the newly derived parameters.

The SVR algorithm just described is for data that has linear relationship. For non-linear data, kernel is used to map the linear features into a higher non linear hyperplane. The idea of a kernel is explained in the next section below.

SVR kernel

A kernel is a function that is used to transform linear SVR algorithm into non-linear by mapping the input value into a higher dimension. The idea of a kernel is that for a non-linear situation, if the data is modified, then it might be possible to linearly separate the data.

There are different kernel functions with support vectors, the popular ones are;

- Polynomial kernel

○ $K(x, y)=(1+x.y)^s$ where s is the degree of the polynomial; x and y represent the input and the output respectively.

- Radial basis function(RBF)

○ $K(x, y) = \exp(-(x-y)^2/2\sigma^2)$ this kernel is also know as gausian kernel and σ represent the variance. This is the kernel function that will be used in this project.

- Sigmoid function kernel

○ $K(x,y)= \tanh(\kappa x.y- \delta)$

Forecasting methodology

Data source

The data used for the analysis and modeling is a time series dataset, which was obtained from the Ministry of Health in Ibadan Region of Oyo State. It was a secondary data. The data on monthly basis consist of the measles cases from various Local governments in Oyo state for the period of January 2008 to June 2015. The model used to analyze the collected data is the Support Vector Regression.

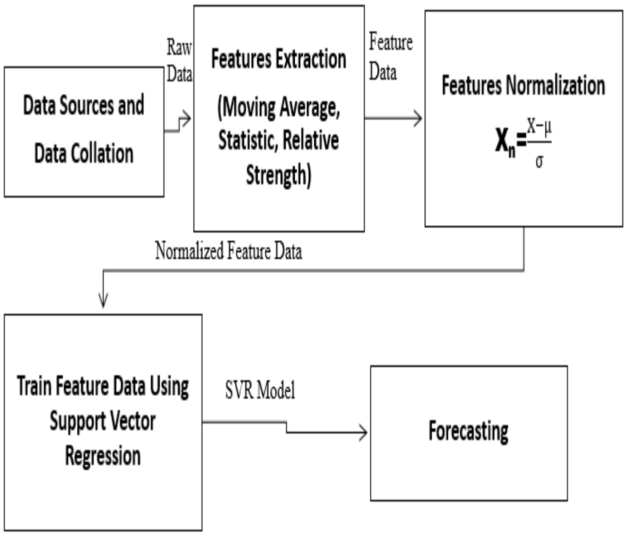

Data Sources and Data Collation
Raw Data
Features Extraction (Moving Average, Statistic, Relative Strength)
Feature Data
Features Normalization
$X_n = \frac{X-\mu}{\sigma}$
Normalized Feature Data
Train Feature Data Using Support Vector Regression
SVR Model
Forecasting

Figure 1. Methodology Summary

Input and output features with representations

In order to accurately forecast the outbreak of measles, the factors that affect the spread must be considered. Although the data gathered for this project is a monthly record of the number of outbreaks in each local government in Oyo state; this time series data will be transformed into features that represent the factors that could affect measles outbreak. The extracted features will then be used as the input values to the SVR training algorithm. The features are listed and explained below;

- *Moving Average*

Moving average is one of the popular parameters for time series data estimation.The moving average for a particular sample data is calculated by getting the previous t-1 and the current data. For example, if the value of t = 3,the previous two values and the current value will be added together and the result will be divided by t. It can be represented mathematically as:

$$M_t = \frac{Xt + Xt-1 + \cdots . + Xt-n+1}{n} \quad (6)$$

Where n = t =3(the value used in this work).

- *Relative strenght indicator*

Relative strength indicator is a measure of upward and download movement. It helps to measure the degree trends and the degree of flunctuation of the time series data. It is the ratio of the number of upward movement and downward movement within a specific time. It can be mathematically represented as:

$$\mathbf{S_t} = \frac{\boldsymbol{Ut}}{\boldsymbol{Dt}} \quad (7)$$

where Ut and Dt represent the upward and downward movement respectively**.** The strenght indicator is calculated for a period of 4 months. That is the value of t=4 since the data is a monthly time series.

- *Statistic*

This feature is calculated by taking the number of infected persons at a particular time and subtract the lowest value within the specified time interval and then divide the number by the range (diffence between the highest and the lowest number) within the time interval. it is shown mathematically below

$$P(t) = \frac{v(t)-L}{H-L} \quad (8)$$

where v(t) is the value at time t, L is the lowest value within the time interval and H is the highest value within the time interval. 4 will be used as the interval value.

- *Output feature*

The output feature will be the number of outbreak at a specific time.The output value will then be used to determine the occurrence of measles outbreak by setting a threshold value of 4. This implies that when the output value is greater than 4 then there is an outbreak, otherwise there is no outbreak of measles.

Input data normalization

Different calculations are performed on the raw data as explained in the previous section to generate each of the features. Consequently, each feature will have a different range of values since the formula used to extract each one feature value is different values in some feature might be relatively higher or lower than that of other features which might cause the learning algorithm to perform poorly. In order to fix this problem, the input data will be normalised. The normalization formula is presented below:

$$X_n = \frac{X-\mu}{\sigma} \qquad (9)$$

X_n is the normalised value, X is the raw value, μ is the mean value for the particular feature, and σ is the variance.

In order to perform normalization, feature values must have been generated for each row of the raw data. The feature value data generated is then normalized to improve the performance of the training." μ" is calculated by getting the average value of all the feature values for a particular feature. The mean is calculated using this formula below:

$$\mu = \frac{\sum x}{n} \qquad (10)$$

Where x is the value for a particular feature and n is the number of those values present for the features.

σ is calculated by using the standard deviation formula as shown below:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}} \qquad (11)$$

Where x is the feature value, μ is the mean and n is the number of values for the features.

The value of n is the same for all the features since the value for all is generated for the same raw data.

Result and Discussion

Figure 2 shows the raw data displayed graphically. The graphical representation is labeled with the local government that owns the measles data.

There is a parameter used for setting how the features were being extracted. The parameter needed for extraction here is the **windows size** and it represents the number of previous data that will be selected in order to estimate each feature from the measles record data. It allows the selection of local government whose features are to be extracted.

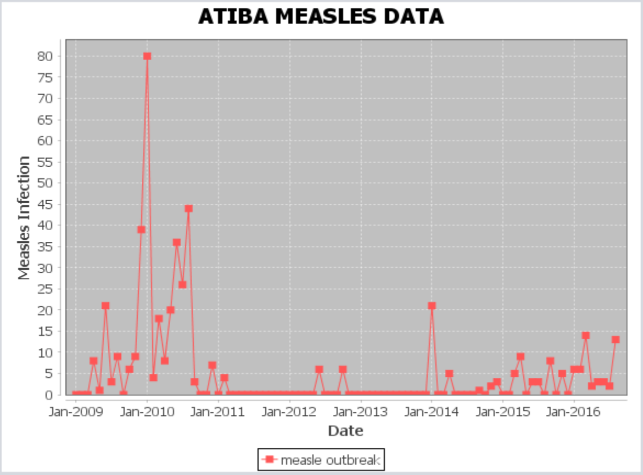


Figure 2: Information displayed by the data page when Atiba Local government is selected from the drop down button.

The maximum value that can be set for the window size is a year (i.e., 12 months) and the higher the window size value, the less the number of training data.

The forecasting and training phase allows the user to set the parameters needed for the training algorithm. Default parameters are provided (except for the local government selection) to enable quick testing even without the knowledge of what the parameter does.

The Information about the parameters are explained below

- C: the penalty value that determines how much the algorithm is penalized for getting the wrong value during the training process
- Epsilon **ε**: this is the minimum error value that must be attained before the training stops. It is the threshold that determines when the training iteration will stop.
- P: this is the value that enables the algorithm to perform soft-margin training and it is related with the epsilon value explained above. From the epsilon information, when the error value obtained after training is subtracted from epsilon the value must be equal to zero(0), however when the p value is set, the algorithm can stop when the error value obtained plus the p value is equal to the epsilon.
- Gamma: The value is needed for kernel value calculation. It is the σ used in the mathematical formula for RBF kernel in the previous chapter.

A proper measure is taken that unsets the parameter values in all fields so that a new value can be added.

The SVR forecasting has an enabling feature which controls the possibility of determining when data training has been performed. Successful data training generates a model which is dependent on which of the local government data was trained. The forecast is done using the recently generated model from the last training. Also, the user can select the desired model from the list of local government as applied.

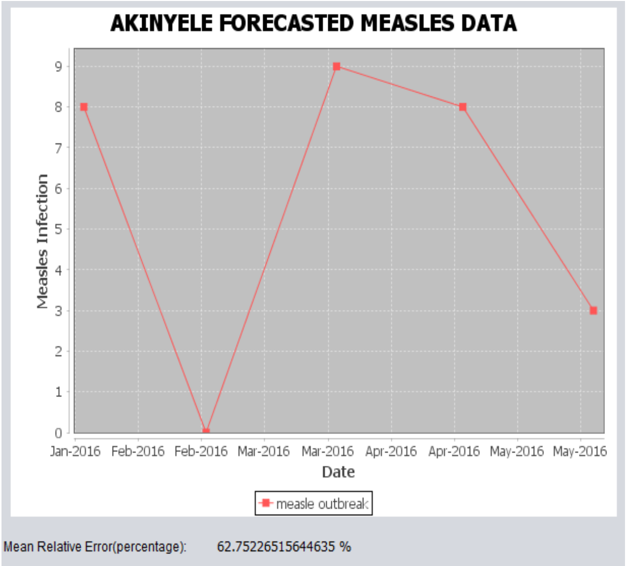
The forecast range parameter is the number of months to forecast for and the threshold value is the threshold value that determines if there is an outbreak or not.

Result Presentation

The dataset for Akinyele local government was trained using the following parameters; C = 100, Epsilon ε = 0.0001, P = 0.001, window size = 3 and gamma = 100.

The trained data was selected for the forecast to be done using a forecasting range of 5 and a threshold value of 5.

The result of the forecast is as shown in figure 3 below. With a threshold of 5, every point on the graph that corresponds to values greater-than or equal to 5 marks a positive reading of the outbreak, otherwise, there is no outbreak.



Mean Relative Error(percentage): 62.75226515644635 %

Figure 3. Akinyele measles forecasting result with MRE of 62.8%.

The result is also presented in a tabular form which shows the serial number, date, the predicted number of infection and outbreak size (used to determine if there is an outbreak or not based on the threshold value set from the training and forecasting phase). The algorithm performance values that were put to use are the mean squared error value and the mean relative error. The explanation for each of the value and how they are derived are explained below.

Mean Squared Error (MSE): The mean squared error is calculated using the mathematical formula below:

$$MSE = \sum_{n}^{1} \frac{(previousValue - predictedValue)2}{n} \qquad 12$$

Predicted Value is the value generated from the learning algorithm after training. Previous Value is the value from the training data. Each value is the value from the corresponding month of the previous year data; n is the number of forecast value from training and prediction phase that generated the result.

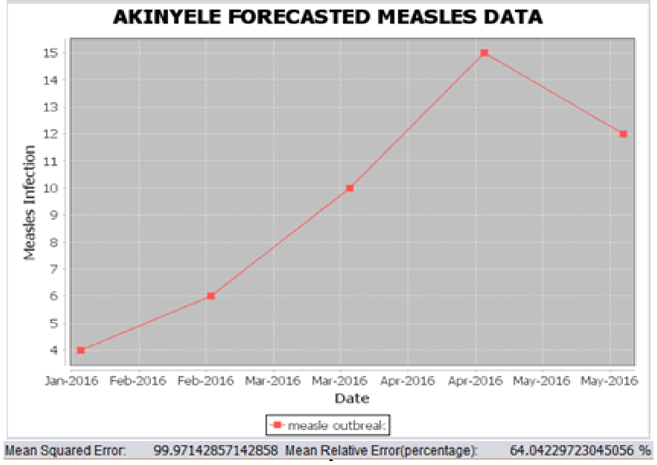
Mean relative error (MRE): This metric measures the percentage closeness of the predicted value to the corresponding month value of the previous year in the training data. The value is estimated using the mathematical formula below:

$$MRE = \sum_{n}^{1} \frac{(previousValue - predictedValue) \times 100}{previous} \qquad (13)$$

It measures the closeness ratio of the predicted Value to the previous Value.

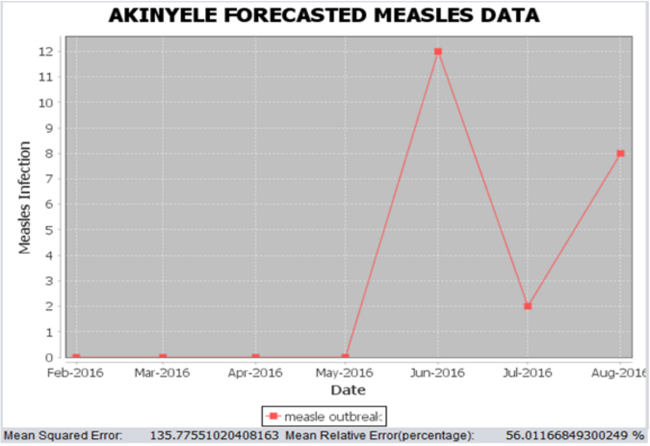
NB: The previous Value and predicted Value in this metric hold the same meaning with that of the previous metric.

Another parameter of significance is the **window size**. The windows size affects the amount of previous knowledge incorporated into the features generated for the data training operation.



Mean Squared Error: 99.97142857142858 Mean Relative Error(percentage): 64.04229723045056 %

Figure 4. Result output when window size is set to 6 while carrying out Measles outbreak forecast for Akinleye local government.



Mean Squared Error: 135.77551020408163 Mean Relative Error(percentage): 56.01166849300249 %

Figure 5. Result output when window size is set to 12 while carrying out Measles outbreak forecast for Akinleye local government.

Figures 4 and 5 compare the output generated when the window size value is set to 6 and when it is set to 12 and having all other parameters constant. The data for Akinyele local government was trained under constant parameters, and varying the window size.

Result discussion

It took 385 milliseconds to complete training when when the window size value is set to 6, and takes 425 milliseconds when the window size value is set to 12.

As shown in figures 4 and 5, Window size value also affects the generated forecast result. If we compare the output in figure 4 and 5 it can be seen that the efficiency of the system increases as window size value increases. When the window size value is set to 6, the mean error is approximately 99.97 and the mean relative error is approximately 62.04%. However, when the window size value is set to 12, the mean error is approximately 135.78 and the relative mean error is approximately 56.01%. Therefore, it can be concluded that the value of the window size set affects the training time of the algorithm and the efficiency of the model generated.

Conclusion

This research did not consider modes of vaccination and modes of treatment as methods to prevent the outbreak of measles in the thirty-three (33) local government regions in Oyo state. It concentrated on the monthly time frame and monthly outbreak size of the herd. Results from this project show that increasing the windows size affects the amount of previous knowledge incorporated into the features generated for the data training operation. Therefore, it can be concluded that the value of the window size set affects the training time of the algorithm and the efficiency of the model generated. That is, as the window size value increases, both the training time and the efficiency of generated models increase in view of the outbreak threshold.

References

[1] Bradley, P. 2012. Predictive analytics can support the ACO model. *Journal of the Healthcare Financial Management Association*, 66, (4), pp.102-106.

[2] Myers M.F., Rogers D.J, Cox J, Flahault A., and Hay S.I. 2000. Forecasting Disease Risk for Increased Epidemic Preparedness in Public Health. *Adv Parasitol* . 47:309-330

[3] James W. Buehler,Richard S. Hopkins, J. Marc Overhage, Daniel M. Sosin,Van Tong. 2004 . Framework for evaluating public health Surveillance systems for early detection of outbreaks; recommendations from the CDC working group. *MMWR*;53:5. pp1-11.

[4] Pinner, R., Rebmann, C., Schuchat, A., and Hughes, J. 2014.Disease Surveillance and the Academic, Clinical, and Public Health Communities. *Emerging Infectious Disease*s, 9(7), pp781–787.

[5] Kandananond, Karin. 2012.A Comparison of Various Forecasting Methods for Autocorrelated Time Series. *International Journal of Engineering Business Management.* 4(4). pp(). 10.5772/51088.

[6] H Chan, Emily & Scales, David & Brewer, Timothy & C Madoff,Lawrence & P Pollack, Marjorie & Hoen, Anne & Choden, Tenzin & Brownstein, John. 2013. Forecasting High-Priority Infectious Disease Surveillance Regions:A Socioeconomic Model. Clinical infectious diseases .*The Infectious Diseases Society of America*. 56(4):pp517-24

[7] Roberts M. G. and Tobias M. I. 2000. Predicting and Preventing Measles Epidemics in New Zealand: Application of a Mathematical Model. *Epidemiol infct* 124, pp279 – 287

[8] K. Y. Chen and C. H. Wang.2007. A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. *Expert Systems with Applications.* 32,(1), pp. 254–264.

[9] AD Cliff and P. Haggett. 1993. Statistical modelling of measles and influenza outbreaks. *Journal of {Statistical Methods in Medical Research.* 2 (1) pp 43-73

[10] R. C Souza. 1992.Forecasting the Progress of Epidemics by Means of a Bayesian-Entropy Framework,Environment and Planning A: *Economy and Space.* 14(1) pp. 49 – 60

[12] Cryer D. Jonathan and Kung – Sik Chan .2008. Time Series Analysis with Applications in R, Second Edition, U.S.A., Springer.501pp

[13] V. N. Vapnik. (1999). The nature of statistical learningtheory. New York: Springer.314pp

Journal of Science Research, ISSN 1119 7333
Citation: Akomolafe O.P., Adewuyi P. and Nzenwata U.J.
Forecasting infectious disease outbreak using Support Vector Regression (SVR) case study: measles (rubeola). Volume 16, 2017, pp7-13