Predicting Floods in North Central Province of Sri Lanka using Machine Learning and Data Mining Methods

H. Thilakarathne[1], K. Premachandra[2]
[1]Department of Physical Science, Faculty of Applied Sciences, Rajarata University of Sri Lanka, Mihintale, Sri Lanka.
[2]Center for Micro-Photonics, Department of Physics and Astronomy, Swinburne University of Technology, Melbourne, Australia.

Abstract

The frequency of the occurrence of natural disasters at present has increased due to changes in global and regional climate. Hence, being able to forecast natural disasters has shown to be extremely useful in mitigating the loss of damages to the mankind. In this study, a hybrid model is developed for predicting the occurrence of floods in the North Central Province of Sri Lanka, using machine-learning techniques with artificial neural networks. The hybrid model developed combines two sub predictive models. First model predicts the future weather-related measurements using time series modeling. The second model, which is a binary classification machine-learning algorithm, predicts the probability of the occurrence of flood incidences in a future month using the forecasted weather values and historical flood data. The results show that all probabilities predicted by the hybrid model are a 91.7% match to the actual flood occurrences. Hence, this model could be adopted to predict flood occurrences of any region in Sri Lanka using historical weather and flood related data of the region of interest. The predictive model developed has been published as an Application Programmable Interface on Microsoft Azure cloud, illustrating the practical usage and feasibility of machine learning techniques in developing modern intelligent applications.

Keywords: Time Series Modelling, Binary Classification Machine-Learning Algorithm, Predict Flood Occurrences in Sri Lanka.

1. Introduction

Natural disasters such as floods, Tsunamis, drought, epidemic diseases, create major environmental economic and political issues, while causing great danger to human lives in Sri Lanka. Floods in Sri Lanka occur mainly during the monsoon season in which major floods in North Central province usually occur during the Northeast monsoons (December to February). Since preventing a natural disaster is not a practical task, predicting them has become extremely helpful for the policy makers and authorities to alert the public and take necessary precautions to avoid them.

In recent years, nonlinear modelling capability of Artificial Neural networks (ANNs) has become widely used in developing nonlinear predictive models for weather analysis [1]. ANNs, which is a subset of machine learning (past experience is used to optimize a performance criterion) [2], is a method that is inspired by the structure and functional characteristics of biological neural networks. ANNs are used in prediction, clustering, and classification (categorize observations). ANNs contain three main interconnected components: input neurons, output neurons and hidden neurons. Hidden neurons are the elements in between input and output neuron layers. The main feature of neural networks is the iterative learning process in which training data cases are presented to the network one at a time while adjusting the weights associated with the input values. After all cases are presented, the process often starts over again. During the learning stage, the network learns to predict the correct output label of training samples by adjusting the weights. A training dataset in machine learning is a set of data used to identify potential relationships between data. A test set is the set of data used to assess the effectiveness of a predictive relationship [3, 4].

ARIMA model is a statistical model that can be applied for the analysis of time series data. It is a generalisation of an autoregressive moving average model that is used to predict future points in a time series.

In this paper, a computational model using ANNs is developed to predict the probability of the flood incidents occur in the North Central Province of Sri Lanka by investigating the

underlying mechanisms of the flood occurrences. The ANNs in this study have been trained with Levenberg-Marquardt backpropagation algorithm, which is specially designed to solve non-linear least square complications [2]. In addition, ARIMA techniques used in the studies where the predictive models are developed are adopted to perform the time series predictions [5].

The forecasting models are developed in Microsoft Azure Machine Learning Studio, a cloud based machine-learning toolkit, which enabled to analyze and process a large volume of data efficiently [6].

2. Literature Review

Meteorological forecasting by means of numerical models dates back to the early 19th century when a mathematical approach for forecasting was proposed by Abbe [7] in the paper, *"The physical basis on long range weather forecasting"*. However, the numerical forecasting is not so accurate since the scientists lack the knowledge in simplifying complex atmospheric dynamics (occur due to variations of weather) into simple mathematical equations. Since then the nonlinear characteristics of weather data have kindled the interest in scientists to use nonlinear prediction mechanisms for both weather and disaster forecasting.

A. *Machine learning models for weather and disaster prediction*

With the evolution of computers, developing computational models for the purpose of weather forecasting became more accurate. In the paper on the origins of computer weather prediction and climate modeling, Lynch was able to show this by describing the evaluation of computer weather prediction and the methodologies used in the particular domain. He further describes the problems that are prevailing in the numerical weather prediction such as unavailability of linear simplified mathematical equations for weather and climate related parameters [8].

The study on "*Weather forecasting model using Artificial Neural Networks"* is carried out to determine the applicability of ANN approach for developing nonlinear predictive models in weather forecasting [1]. The advantages of having ANNs for weather forecasting over other forecasting methods are emphasized in this study. Since ANNs minimize the errors using various algorithms, the performance is improved in contrast to other models in the weather prediction domain. The tool used to carry out the analysis is Neural Network Fitting Tool, *nntool* available in MATLAB software. Artificial Feed-Forward Neural Network with back-propagation principles is selected as the training element.

B. *Time series predictions for weather forecasting*

Autoregressive Integrated Moving Average (ARIMA) techniques have been used to predict the behavioural pattern of weather attributes such as rainfall in the literature [5, 9, 10]. ARIMA is a statistical technique for modeling time series data [4]. To maximize the prediction accuracy of ARIMA models, model selection is performed over a time series in an automated fashion [5].

The numerous weather and disaster prediction studies provide an extensive base of information and approaches to the problem. These studies cover a wide spectrum of computational efforts from traditional statistical approaches such as the numerical weather prediction modeling, to machine learning and data mining methodologies [7, 8, 11].

In our study on building a model for predicting floods in North Central province of Sri Lanka using machine learning and data mining methods, artificial neural networks are used to adopt the non-linearity in predictive models. ARIMA models are adopted to perform the time series predictions within the model.

3. Methodology

A. *Data Collection*

Weather related historical data (January 1976 to December 2015) of Anuradhapura district in the North Central Province of Sri Lanka including monthly average rainfall data (mm), monthly average minimum and maximum temperature data (°C) have been collected from the Department of Meteorology of Sri Lanka (Table 1) [12]. Flood related historical data of the North Central Province of Sri Lanka is extracted from *DesInventar* – Disaster Information Management System of Sri Lanka [13]. The data includes 51 flood type disaster records from January 1976 to December 2015.

Table1. Statistical summary of the historical weather data obtained from the department of meteorology, Sri Lanka.

	Weather Attributes		
	Rainfall	**Minimum Temperature**	**Maximum Temperature**
Number of records	480	480	480
Mean	111.71mm	23.80°C	32.73°C
Median	73.25mm	24.10°C	33.05°C
Minimum Value	0mm	18.80°C	28.30°C
Maximum Value	683.9mm	26.10°C	37.50°C
Standard Deviation	120.78	1.38	1.80
Number of Unique Values	399	89	100

All data are converted into a monthly attribute and the missing values are substituted by the attribute's mean (average) of the specific field since it is independent of the sample mean [14].

B. Predictive modeling

In order to build the flood type disaster prediction model, two sub predictive models are developed in the study. While the first model involves predicting future values of the weather attributes (average monthly rainfall, average monthly maximum and minimum temperatures), the second model involves forecasting the flood type disaster incidents based on the input weather parameters. The output values of the first predictive component have been used as the input variables of the second predictive component.

The future weather attributes, which are the output of the first component of the flood prediction model, are forecasted using the ARIMA model. By defining historical data as a continuous time series, monthly weather data over a span of 39 years (January 1976 to December 2015) have been used as input variables in this component of the model. The values of weather attributes have been forecasted for the period of one year from January 2015 to December 2015. These predicted values are then compared to the actual weather data from January 2015 to December 2015 to evaluate the accuracy and the reliability of the developed time series model.

The model that predicts the probability of flood incidents is developed as the second component of the hybrid model. It is identified as a binary classification machine-learning problem, where the instances are classified into two classes [2].

The forecasted weather values are utilized as input data of this component of the model. The input data parameters are normalized to values between 0-1, using *minmax normalization* [15]. The output of the model is calculated as a probability, where the values approximate to 1 denoting a positive flood type disaster occurrence and 0, a negative (no) flood type disaster occurrence in the specific month to the future. The data flow diagram of the complete predictive model is shown in figure 1.

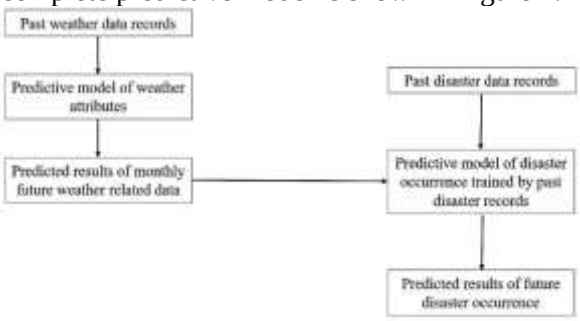

Figure 1: Data flow diagram of the complete predictive model

A binary classifier built with a neural network having a learning rate of 0.70, two hidden neurons and one hidden layer is chosen to build the predictive experiment. The optimal learning rate and the number of iterations are chosen using the '*Tune Model Hyperparameters*' module in Azure Machine Learning Studio [16]. This model uses different combinations of settings of the neural network in order to determine the optimum hyperparameters (parameters whose values are set prior to the commencement of the learning process) for the given prediction task and data [16]. The learning rate is a systematic approach to generalize the model when a large amount of training data is used. The classification model using the training dataset is cross-validated for reliability by dividing the dataset into 10 folds and calculating the classification accuracy for each fold [14].

3. Results

A. Output of the 1st predictive model

The predicted output values of rainfall, minimum temperature and maximum temperature of year 2015, show a slight deviation between the predicted and the actual values (Table 2, Figure 2,3,4). However, it is interesting to notice that both forecasted and actual values of all attributes follow the same distribution pattern (Figure 2,3,4).

The forecasted *accuracy measures* comparing predicted values to the observed values are obtained via calculating the root mean square error (RMSE) of each predicted attribute. The RMSE values of forecasted rainfall, minimum and maximum temperatures (115.58, 0.42 and 0.56) show a relatively high error rate compared to the forecasted minimum and maximum temperature values.

Table2. The actual values of the weather attributes in 2015 vs the predicted output of the first predictive model.

Month	Weather Attribute					
	Rainfall		Minimum Temperature		Maximum Temperature	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
1	15.8	189.79	22.1	21.71	29.7	30.59
2	161.1	154.29	23	22.37	31.3	31.94
3	26	119.20	23.9	23.45	33.6	34.58
4	288.2	206.53	25	24.40	34	34.21
5	264.7	65.46	25.9	25.44	33.4	33.91
6	10.1	26.42	25.5	25.47	33.9	34.00
7	0	32.47	25.3	25.055	34.1	33.99
8	180.2	54.85	25.1	24.98	34.2	34.20
9	147.1	81.70	24.8	24.80	33.5	34.57
10	388.6	257.33	24.6	23.99	32.8	33.11
11	452.4	263.20	23.8	23.26	31.6	31.41
12	286.4	262.17	23	22.82	29.8	29.90

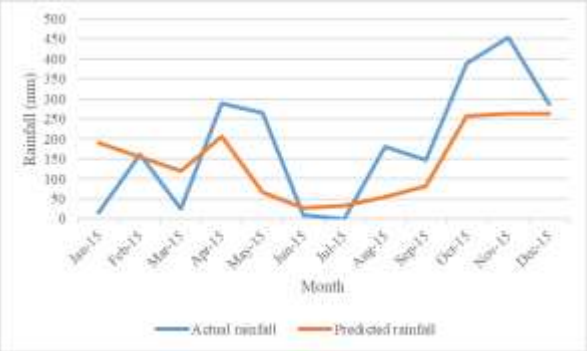

Figure 2: Deviation of actual rainfall data and predicted rainfall values for the year 2015

Figure 3: Deviation of actual minimum temperature data and predicted minimum temperature values for the year 2015

Figure 4: Deviation of actual maximum temperature data and predicted maximum temperature values for the Year 2015

B. Output of the 2nd predictive model

The classification model with 2 hidden neurons and 0.70 learning rate, shows that the model predicts the occurrence of floods with a mean accuracy of 0.917 and a 0.044 standard deviation (Table 3).

Table3. 10-fold cross validation output of the binary classification model.

Fold Number	Accuracy	Precision	Recall	AUC
0	0.854	1.000	0.364	0.958
1	0.917	0.500	0.250	0.747
2	0.875	0.800	0.444	0.923
3	0.958	1.000	0.714	0.951
4	0.958	1.000	0.600	0.944
5	0.979	0.500	1.000	0.979
6	0.917	0.000	0.000	0.815
7	0.917	0.333	0.333	0.911
8	0.938	0.500	0.333	0.889
9	0.854	0.000	0.000	0.948
Mean	0.917	0.563	0.404	0.907
Standard Deviation	0.044	0.384	0.308	0.073

The receiver operating characteristic (ROC) curve plotted between true positive rate and false positive rate (Figure 5) for this two-class classification problem is used to evaluate the accuracy of classification of the flood type disaster prediction model (In a ROC curve the true positive rate is plotted in relation to the false positive rate for different cut-off points of a parameter).

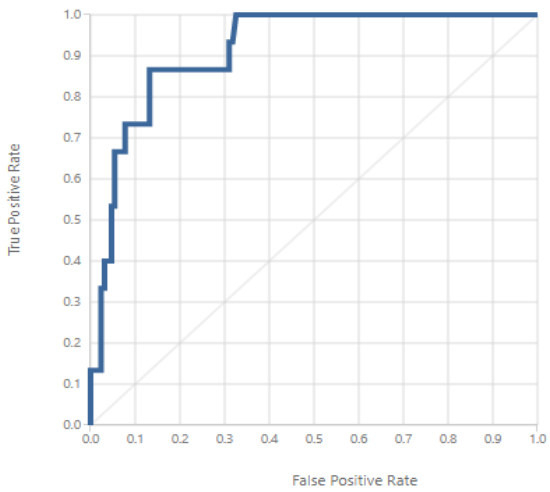

Figure 5: ROC curve of the flood type disaster prediction model

The predictions are yielded to a point in the upper left-hand corner of the ROC space denoting a high probability of a true positive output. The (0,1) point is called a '*perfect classification*', where no false positive or false negative outputs are predicted using the model. The true positive rate defines the number of correct positive results among all positive samples available during the test. False positive rate, on the other hand, defines the number of incorrect positive results among all negative samples available during the experiment [17]. The area under the curve (AUC) is 0.915 square units, showing that the probability of a randomly chosen positive instance is higher than a randomly chosen negative one.

A web application programmable interface (API) is implemented using the flood predictive model in order to obtain the predicted values of flood incidents in the North Central Province of Sri Lanka (https://goo.gl/faICPD).

4. Discussion

Natural disasters cause significant damages to the environment while causing a threat to human lives. The ability to predict disasters prior to occurring helps minimize the damages caused. This study proposes a predictive model with two sub predictive models that forecasts floods using historical weather and flood data. ANNs and statistical concepts of time series forecasting are utilized to develop the model.

RMSE values of time series predictions of weather data (average rainfall, average minimum temperature, and average maximum temperature) obtained from the first predictive model are 115.58, 0.42, and 0.56 respectively. Average minimum and maximum temperature values show very low RMSE values depicting a very small deviation between the forecasted and actual values. However, high RMSE value of average rainfall data shows a large deviation between the forecasted and actual values. High RMSE of predicted rainfall may have occurred due to outliers. These outliers could be due to the rain gauge measurement errors of the ground meteorology stations where the data is collected. In addition, sudden weather changes such as thunderstorms and hurricanes could have caused outliers in monthly average rainfall data. However, removing outliers from the training dataset could reduce RMSE values and hence, could obtain a better-forecasted output and a reliable predictive model.

Limiting the number of hidden neurons to two neurons in the second sub predictive model minimizes the probability of overfitting. Overfitting occurs when the model becomes an exact fit to the training dataset. This makes the model less generalised and leads to poor performance on new data [2].

The results of the predicted model indicate that the forecasted occurrence of flood incidents accurately corresponds to the testing dataset (Table 3). The binary classification that has been adopted for the model uses 0.5 as its threshold giving the optimal AUC.

Increasing the number of input parameters of the neural network increases the accuracy and the reliability of the predictions [1]. Hence, the model could be more reliable if weather data such as relative humidity, atmospheric pressure, wind speed and solar radiation are taken into consideration.

5. Conclusions

Being able to predict natural disasters has become an extremely important task to the policy makers and authorities since they cause significant damage to the environment and human lives. In this paper, a computational model using machine learning techniques with ANNs has been proposed to predict the occurrence of flood incidents in a particular month in the North Central Province of Sri Lanka.

This predictive model is built as a combination of two separate forecasting models. The first model forecasts the future values of weather attributes using time series forecasting methods (ARIMA). The second model predicts the probability of flood occurrences in a

particular month as a two-class classification machine-learning task.

The results of the predictive model show that it performs with a 91.7% accuracy, proving that this model can be used to predict the occurrence of flood incidents in any region in Sri Lanka using the historical data on weather and floods of the region of interest.

This study also proves that machine-learning models interconnected with data mining approaches are capable of performing environmental predictions based on historical data.

References

[1] K. Abhishek, P. M. Singh, S. Ghosh and A. Anand, "Weather forecasting model using Artificial Neural Network," Procedia Technology, vol. 4, pp. 311-318, 2012.

[2] A. Smola and S. Vishwanathan, Introduction to Machine Learning, Cambridge University Press, 2010.

[3] I. Bose and R. K. Mahapatra, "Business Data Mining - a machine learning perspective," Information & Management, pp. 211-225, 2001.

[4] I. Ibragimov, "Encyclopedia of Mathematics," 2 February 2011. [Online]. Available: http://www.encyclopediaofmath.org/index.php?title=Time_series&oldid=16499. [Accessed 14 August 2016].

[5] M. J. Kane, N. Price, M. Scotch and P. Rabinowitz, "Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks," BMC Bioinformatics, vol. 15, no. 276, 2014.

[6] J. Barnes, Azure Machine Learning, Redmond, Washington: Microsoft Press, 2015.

[7] C. Abbe, "The physical basis of long range weather forcasting," Mon. Weather Rev., vol. 29, pp. 551-561, 1901.

[8] P. Lynch, "The origins of computer weather prediction and climate modeling," Journal of Computational Physics , 2008.

[9] S. M. Sundaram and M. Lakshmi, "Rainfall Prediction using Seasonal Auto Regressive Integrated Moving Average model," Indian Journal of Research, vol. III, no. 4, pp. 58-60, 2014.

[10] D. K. Patrick, P. P. Edmond, T. M. Jean-Marie, E. E. Louis and K.-t. N. Ngbolua, "Prediction of rainfall using autoregressive intergrated moving average model: Case of Kinshasa city (Democratic Republic of the Congo), from the period of 1970 to 2009," Journal of Computation in Biosciences and Engineering, vol. 2, no. 1, 2014.

[11] D. P. Solomatine and Y. Xue, "M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China".

[12] "Department of Meteorology Sri Lanka," 2016. [Online]. Available: http://www.meteo.gov.lk. [Accessed 1 October 2017].

[13] "Disaster Information Management System - Sri Lanka," Disaster Management Centre (DMC),Ministry of Disaster Management, [Online]. Available: http://www.desinventar.lk. [Accessed 6 August 2016].

[14] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2012.

[15] G. K. Patro and K. K. sahu, "Normalization: A Preprocessing Stage," in arXiv preprint arXiv:1503.06462 , 2015.

[16] Microsoft, "Tune Model Hyperparameters," 21 July 2017. [Online]. Available: https://msdn.microsoft.com/en-us/library/azure/dn905810.aspx. [Accessed 15 September 2017].

[17] L. G. Michael Goebel, "A Survey of Data Mining and Knowledge Discovery Software Tools".

[18] M. S. G. A. A. Kumar Abhishek, "Weather forecasting model using Artificial Neural Network," Procedia Technology , pp. 311-318, 2012.

[19] P. Baldi and S. Brunak, Bioinformatics: The Machine Learning Approach, The MIT Press, 2001.