MALARIA INCIDENCE FORECASTING FROM INCIDENCE RECORD AND WEATHER PATTERN USING GMDH POLYNOMIAL NEURAL NETWORK

Anditya Arifianto[1], Ari Moesriami Barmawi[2], Agung Toto Wibowo[3]

[1]Magister Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Malaria affects over 100 million persons worldwide each year. The impact of malaria can cause approximately 2,414 deaths a day in average. Indonesia has a great number of incidents it is on the third highest position of malaria incident in South East Asia, with number of confirmed cases of 229,819 cases reported and 432 deaths only at 2010. The Malaria incidence prediction is badly needed so that the Health Department of Ministry of Indonesia is able to make the necessary preparation to prevent and reduce the impacts. Malaria incidence Prediction is a problem of Time series prediction, and a Time series prediction involves the determination of an appropriate model, which can encapsulate the dynamics of the system, described by the sample data. Previous work has demonstrated the potential of neural networks in predicting the behavior of complex, non-linear systems. Group Method of Data Handling (GMDH) Polynomial Neural Network was applied in a great variety of areas for data mining and knowledge discovery, forecasting and systems modeling, optimization and pattern recognition. This paper proposed a modified GMDH Polynomial Neural Network to reduce the learning time and computation while maintaining the accuracy. Based on the experiments, it was proven that the modified GMDH PNN was able to reduce the learning time by 72% and improve the accuracy by 7%, 5.37%, and 4.97% into the accuracy of 88.02%, 86.12%, and 83.90% for 1st, 2nd, and 3rd month prediction compared to the original GMDH PNN.br>

Kata Kunci : Malaria, Prediction, Polynomial Neural Network.

Abstract

Malaria affects over 100 million persons worldwide each year. The impact of malaria can cause approximately 2,414 deaths a day in average. Indonesia has a great number of incidents it is on the third highest position of malaria incident in South East Asia, with number of confirmed cases of 229,819 cases reported and 432 deaths only at 2010. The Malaria incidence prediction is badly needed so that the Health Department of Ministry of Indonesia is able to make the necessary preparation to prevent and reduce the impacts. Malaria incidence Prediction is a problem of Time series prediction, and a Time series prediction involves the determination of an appropriate model, which can encapsulate the dynamics of the system, described by the sample data. Previous work has demonstrated the potential of neural networks in predicting the behavior of complex, non-linear systems. Group Method of Data Handling (GMDH) Polynomial Neural Network was applied in a great variety of areas for data mining and knowledge discovery, forecasting and systems modeling, optimization and pattern recognition. This paper proposed a modified GMDH Polynomial Neural Network to reduce the learning time and computation while maintaining the accuracy. Based on the experiments, it was proven that the modified GMDH PNN was able to reduce the learning time by 72% and improve the accuracy by 7%, 5.37%, and 4.97% into the accuracy of 88.02%, 86.12%, and 83.90% for 1st, 2nd, and 3rd month prediction compared to the original GMDH PNN.

Keywords : Malaria, Prediction, Polynomial Neural Network.

CHAPTER I
INTRODUCTION

1.1 Rationale

Malaria affects over 100 million persons worldwide each year. The impact of malaria can cause about 2,414 deaths a day. It is both a disease of poverty and a cause of poverty slowing economic growth by 1.3% per year in endemic areas [**1**]. Malaria is an enormous health and development problem in the South East Asia (SEA) Region as 1,322 million people of South East Asia Region or 76% of the total population is at risk of malaria. Among the population living in malaria endemic areas, infants, young children and pregnant women have been identified as high risk groups [**2**].

In 2000-2010, around SEA Region, it was reported that malaria incidence remained around 2.30 – 3.08 million cases and deaths around 2,423 – 6,978 annually. During 2010, the confirmed malaria cases and malaria deaths were reported as 2.3 million and 2,426 respectively in the Region, whereas the WHO/HQ estimated malaria cases and deaths were around 28-41 million and 49,000 respectively [**2**]. Indonesia located in SEA has a great number of incidents, the third highest position of malaria incident, with number of confirmed cases of 229,819 cases reported and 432 deaths [**3**].

Due to the severe health impact of malaria epidemics there is a growing need for methods that allow forecasting, early warning and timely case detection in areas of unstable transmission, such as the Indonesian inland, so that more effective control measures can be implemented. Studies of malaria epidemics in these areas have shown their association with excess rainfall, temperature and vegetation density measured by the normalized difference vegetation index (NDVI) and also the *Anopheles* mosquito as the medium. [**1**,**4**]

The *Anopheles* mosquitos like to live in areas that have an abundance of natural water because they need this natural water to breed. You can expect to find these mosquitoes and their larvae in ponds, marches, swamps, ditches, rain pools, and on the shores of streams and rivers. Some breed in shady areas such as forests, while others breed in open fields where there is plenty of sunlight. This is seen in the direct correlation between an abundance of *Anopheles* mosquitoes and rainfall, increased transmission and temperature, and vegetation density and malaria seasonality [**1**,**4**].

Banggai regency, Central Sulawesi, is one of the malaria prone areas because of its agricultural areas, rice fields and marshes. 16 out of 19 sub-districts are at high risk with number of malaria transmission above 5 cases per 1000 population per year.

Malaria incidence forecasting is a problem of time series prediction, and a Time series prediction involves the determination of an appropriate model, which can encapsulate the dynamics of the system, described by the sample data. Previous work has demonstrated the potential of neural networks in predicting the behavior of complex, non-linear systems. In particular, the class of polynomial neural networks has been shown to possess universal approximation properties, while ensuring robustness to noise and missing data, good generalization and rapid learning [**5**].

Group Method of Data Handling (GMDH) Polynomial Neural Network has been applied in a great variety of areas for data mining and knowledge discovery, forecasting and systems modeling, optimization and pattern recognition. Inductive GMDH algorithms give possibility to find automatically interrelations in data, to select optimal structure of model or network and to increase the accuracy of existing algorithms. And Because of that, for this research, to make a prediction system of malaria incidence, a self-organizing neural network known as GMDH Polynomial Neural Network will be used [**6**].

1.2 Theoretical Framework

The problems mentioned before are related to correlate Malaria Incidence and Weather Pattern. The input for this system is time series data of weather factors while the output is the number of predicted malaria incidence in a particular predicted month.

In building system that is able to perform a good prediction, it is necessary to learn some theories and concept related to data reparation, Information Criterion Selection function, feature extraction and prediction method. In the learning process, the prediction model is generated using GMDH PNN with various information Criterion Selection function and cross validation in order to prevent over fitting. The performance measured by testing using testing dataset and the model with the highest accuracy will be used as the main model for the system.

1.3 Conceptual Framework / Paradigm

To achieve the optimal objectives, there are several processes to be observed, starting from data collecting, data preprocessing, and main process in Prediction, until obtaining the output of the system. The following are some variables that need to be paid attention for relating the experiments to general problems in Malaria Prediction.

Table 1-1 : Variable Observation

Concept	Variable	Unit analysis
Composition of data used as training dataset, validation dataset, and testing dataset.	Data composition	Ratio to divide the data into 3 datasets in annual unit. The unit of measurement to analyze is the number of weather factors included in dataset.

This indicating the length of past series data used as input.	Number of time series	The unit analyzed is the number of series data input to the system measured by MAPE of the prediction.
The function used to calculate the output in each neuron.	Polynomial Function	The unit analyzed is the polynomial function used in the learning process measured by MAPE of the prediction.
The parameter used to select better neurons to survive in each layer performed by Information Criterion Functions	Criterion function	The unit analyzed is the criterion function used in the learning process measured by MAPE of the prediction.

1.4 Statement of the Problem

Based on the rationale above, it can be identified that Malaria is a serious disease in Banggai regency. A system to predict the Malaria outbreak is needed so that the Ministry of Indonesia can make a precaution for the impact. However, the Health Department of Ministry of Indonesia has not got a forecasting method with a good performance.

The existing forecasting method owned by the ministry is limited to study of auto regression such as ARIMA. Most standard ARIMA based programs as a linear model for forecasting has a significant weakness to outlier detection. That is, when an outlier occurs not alone do the parameters have to be re-estimated but it may also be the case that order of the ARIMA model has also changed at that point. With the urgency of the problem which has a time series properties, a more adaptive method is needed as the auto regression technique has a limited adaptation toward a further prediction.

Many researches about time series prediction only use one property as input and output. However, to correlate the malaria incidence and weather factors, the system is faced with multiple data series as input. Although GMDH PNN was considered has a fast learning time with its self-organizing properties, the standard GMDH PNN still perform excessive calculation and complexity by blindly try all possible architecture in self-organizing training process. Thus, a modification to reduce the complexity and to guide the learning process is needed so the learning will proceed faster while maintaining its accuracy.

1.5 Objectives and Hypotheses

The first objective of this work was to study the relation among the weather factors and malaria incidence to develop a predictive model that was able to forecast the incidence of malaria with reasonable reliability using the reported case rate, rainfall, temperature, and precipitation day. The system's reliability showed by how high the accuracy of predicted incidence and how much time of upcoming malaria incidence could be predicted.

From the previous researches, The Artificial Neural Network (ANN) proved to be adequate for a malaria forecasting system by having a smaller mean square error and absolute error compared with the logistic regression model and the actual values. It showed some advantages: strong ability for analysis, lower claim for data, convenient and easy to apply. And some of malaria prediction research using ANN showed that prediction of malaria incidence was able to be conducted by connecting the incident with the weather condition.

In this case, GMDH Polynomial Neural Network, as a non-linear approach of ANN, proved to be robust and reliable in term of predicting time series which resulting prediction with a higher accuracy than the linear ANN. As a self-organizing Neural Network, GMDH PNN faster in training and more convenient in adapting low quality data than the linear ANN.

Thus, the hypothesis stated in this work was that Malaria incidence in a particular month can be optimally predicted in term of accuracy and time predicted using GMDH Polynomial Neural Network by means of interconnecting the rainfall, temperature, precipitation day, (as factors that determine the density and infectivity of *Anopheles* mosquitoes) and the malaria incidence data in the preceding month (as an estimator of the human reservoir of the parasite and of population susceptibility) as the input.

The Health Department of Ministry of Indonesia has proposed to use 3 weather factors which are rainfall, temperature, precipitation day as input aside from malaria incidence to make a forecasting system. However, the mosquito's habitat and life cycle are not only affected by those three factors, but also other factors such as humidity and length of daylight time. Therefore, the second hypothesis state in this research is that performance of forecasting system using more weather factors is better than only using the proposed weather factors.

In general form of GMDH PNN with neuron that has 2 inputs, the learning process will try all of C_2^n combinations of possible neuron, with n is number of neuron in preceding layer, to create the network architecture. Thus, the second objective of this work is to make a modification to GMDH Polynomial Neural Network that will reduce the complexity and training time while maintains the prediction performance compared to the original GMDH

PNN. The hypothesis for the second objective is that with a selecting method to supervise the combination trial the learning process will complete faster with better performance

1.6 Assumption

Based on the general and specific problems, some assumptions are developed to support this research. The weather factors data as input for predicting malaria incidence are assumed to be valid data from BPS Statistical of Banggai Regency, and the malaria incidence data assumed to be valid data from Health Department of Banggai Regency which has no missing value.

1.7 Scope and Delimitation

The focus of this study is to use and analyze the methods for predicting malaria incidence based on weather pattern. This study is conducted based on monthly weather data from BPS Banggai and monthly malaria incidence data from Ministry of Health Indonesia from 2004 – 2009.

For the input and output of the system, number of time series of each weather factor is limited from 1 previous month to 5 previous months. The number of predicted Malaria incidence is limited to 3 months predictions. The system only predicts the upcoming malaria incidence without predicting the upcoming weather condition.

1.8 Importance of the Study

The study of the research has some important benefits in the real world. The performance of GMDH Polynomial Neural Network in predicting malaria incidence based on environmental and weather factors can be an alternative method to be used by Health Department of Ministry of Indonesia to predict the upcoming malaria incidence at the regency. By being able to predict the malaria incidence, Indonesian Government, especially Ministry of Health, can make some strategic planning to prevent outbreak and reduce the number of malaria incidence in order to control malaria transmission in Indonesia.

CHAPTER V
CONCLUSION AND RECOMMENDATIONS

This chapter provides the conclusion of the study and recommendations for future work in this area.

5.1 Conclusions

Modified GMDH Polynomial Neural Network, with addition of Entropy and Information Gain to choose the input combination for each neuron in every layer, results a better prediction model than the original GMDH PNN. The modified method was also able to speed the training process by reducing the computational process.

By some observation, the optimal parameter input series were selected for the optimal training process of GMDH PNN. They were 4 to 5 input series using all weather factors data, while trained using polynomial function of Biquadratic and Triquadratic. The optimal criterion function observed was using RMSE Train+Val and using Corrected Akaike's Information Criterion.

The prediction model trained with more weather factors as input performs better than prediction model trained using only three weather factors with accuracy improvement of 9.87% for 1st month prediction, 10.54% for 2nd month prediction and 11.55% for 3rd month prediction. The performance of GMDH PNN overpowered the performance of prediction model trained using ANN BP and ARIMA in term of accuracy.

The implementation of Modified GMDH PNN was able to improve the system performance based on the accuracy rates and quality. The modified GMDH PNN were able to reach prediction accuracy of 88.02% with quality of 0.083 for 1st month prediction, 86.12% with quality of 0.072 for 2nd month prediction, and 83.90% with quality of 0.062 for 3rd month prediction and also reduce the average training time by 72% compared to the original GMDH PNN.

Most of failed predictions were the result of the limited calculation ability of the environment. The system was able to be improved by optimizing the data selection and normalization process.

5.2 Future Works

The treatment method in data preparation will affect the performance of the learning process. Any other normalization and data complexity reduction such as PCA can be applied to improve the prediction performance.

Criterion function is the function that determines the quality of the network and the one that decides whether the learning process needs to be continued or not. Any other quality measurement function and technique may help the learning process to make a better prediction model. Study to minimalize the failed prediction is needed so the performance of GMDH PNN will fully visible.

More data analysis about the connection from weather data and other attributes with malaria incidence was needed to select the data attribute that has the most effect and correlation to the malaria incidence. More additional attributes that may have some contribute to malaria incidence such as NDVI data might improve the prediction result.

BIBLIOGRAPHY

[1] Alberto Gomez-Elipe, Angel Otero, Michel van Herp, and Armando Aguirre-Jaime, "Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1997–2003," *Malaria Journal*, vol. 6, no. 1, p. 129, September 2007.

[2] WHO. (2011, September) WHO Regional Office for South-East Asia. [Online]. www.searo.who.int/en/Section10/Section21/Section340_4018.htm

[3] WHO. (2012, January) Indonesia-Malaria Situation in SEAR Countries. [Online]. www.searo.who.int/en/Section10/Section21/Section340_4022.htm

[4] Suwito, Upik Kesumawati Hadi, Singgih H Sigit, and Supratman Sukowatu, "Hubungan Iklim, Kepadatan Nyamuk Anopheles dan Kejadian Penyakit Malaria," *J. Entomol. Indon*, vol. 7, no. 1, pp. 42-53, April 2010.

[5] Panos Liatsis, Amalia Foka, John Yannis Goulermas, and Lidija Mandic, "Adaptive polynomial neural networks for times series forecasting," in *49th International Symposium ELMAR*, Zadar, Croatia, 2007, pp. 35 - 39.

[6] A.G. Ivakhnenko, "The Review of Problems Solved by Algorithms of the Group Method Data Handling," *Pattern Recognition and Image Analysis*, vol. 5, no. 4, pp. 527-535, 1995.

[7] Eskindir Loha and Bernt Lindtjørn, "Model variations in predicting incidence of Plasmodium falciparum malaria using 1998-2007 morbidity and meteorological data from south Ethiopia," *Malaria Journal*, vol. 9, p. 166, 2010.

[8] R Ghazali, A. J. Hussain, P. Liatsis, and H Tawfik, "The application of ridge polynomial neural network to multi-step ahead financial time series prediction," *Neural Comput & Applic*, vol. 17, pp. 311-33, July 2007.

[9] Sung-Kwun Oh and Witold Pedrycz, "The design of self-organizing Polynomial Neural Networks," *Information Sciences*, vol. 141, pp. 237–258, 2002.

[10] Frank Lemke and Johann-Adolf Müller, "Self-Organizing Data Mining Based on GMDH Principle".

[11] Juan Peralta, German Gutierrez, and Araceli Sanchis, "Design of Artificial Neural Networks Based on Genetic Algorithms to Forecast Time Series," *Innovations in Hybrid Intelligent Systems, Springer*, vol. 44, pp. 231-238, 2007.

[12] Marston B, *The Natural History Of Mosquitos*. New York: The Mac Mollon Co, 1949.

[13] Suyanto, *Artificial Intelligence: Searching, Reasoning, Planning, and Learning. Bandung*.

Bandung: Penerbit Informatika, 2007.

[14] Saeed Farzi, "The Design of Self-Organizing Evolved Polynomial Neural Network Based on Learnable Evolution Model 3," *The International Arab Journal of Information Technology*, vol. 9, no. 2, pp. 124-132, 2012.