VNU Journal of Science: Earth and Environmental Sciences, Vol. 34, No. 4 (2018) 80-88

VNU Journal of Science: Earth and Environmental Sciences

Journal homepage: https://js.vnu.edu.vn/EES

Combination of Adaptive Fuzzy Inference System and Simulated Annealing Algorithm-based for Malaria Susceptibility Mapping in Daknong Province

Bui Quang Thanh*

Faculty of Geography, VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam

Received 23 September 2018
Revised 07 December 2018; Accepted 11 December 2018

Abstract: Adaptive Neuro-Inference system (Anfis) has been widely used in recent studies aiming at generating probabilities of unseen data in binary classification application. It is normally used in combination with optimization algorithms for tuning its parameters to generate optimal objective values. This study proposed a state-of-the-art method using Simulated Annealing to improve Anfis performance. Malaria occurrences and spatial variation of environmental, socio-economic factors in Daknong province, Vietnam were selected for case study. For accuracy assessment, Receiver Operating Characteristic curve, Cost curve were used and the predicted map was compared to several benchmark classifiers. The results showed that the S-Anfis (AUC = 0.912, RMSE =0.335) outperformed Support Vector Machine (AUC = 0.902, RMSE =0.364), Multiple Layer Perceptron (AUC = 0.868, RMSE =0.430). Although, the performance of S-Anfis depended on proper selection of input factors and geographic variations of those, we concluded that this method could be an alternative in mapping susceptibility of malaria.

Keywords: Anfis, Simulated annealing, malaria.

1. Introduction

As report by [1], risk of Plasmodium falciparum (P.f) and Plasmodium vivax (P.v) malaria was significantly worsening in less developed and isolated regions around the world. The most prominent regions are those which have limited accessibility to health services or disease preparedness programs. In which community susceptibility to malaria is one of the key index for disease control and prevention program in every country. Transmission of this disease is mostly influenced by physical environment, climatic and socioeconomic condition.

* Tel.: 84-943672345.
Email: qthanh.bui@gmail.com

https://doi.org/10.25073/2588-1094/vnuees.4304

Currently, the relation of those variables has been studied with support of recent development of spatial technology and data mining techniques. Specifically, susceptible mapping is widely used as it provides probability variations of malaria infection rate as consequence of non-linear modelling of physical and social influential factors. Most recent researches on spatial variation of malaria focused on application of data mining classifiers and their tweeted versions. In which neural network family, support vector machine, decision rules are among common techniques.

Another approach is aiming at exploring natural reasoning with application of fuzzy logics. Fuzzy logic relies on human understanding in defining membership relation between input variables. It is customized to match diversity of input data. Among all fuzzy logic tools, Adaptive Neuro Fuzzy Inference System (Anfis) is one of the most common algorithm in classification application. It is one of the greatest tradeoff among Artificial Neural Networks and fuzzy logic systems. There were many theoretical researches and pratical works aiming at exploring the predictive capability of Anfis, in which the system parameters were tuned by optimization algorithms. There were also several studies on community diseases but few focused on tuning Anfis parameters.

This study proposed a new hybrid method named S-Anfis, using Simulated Annealing optimization algorithm to maximize performance of regular Anfis. Malaria occurrences and independent variables in Dakong provine, Viet Nam were selected as input database for training and validating the proposed model. The rest of the paper is organized as follows: the next section provides description of the study area and data used; the third one introduces research methodology; the fourth includes results and discussions; conclusion and final remarks are in the last section.

2. Data and methods

2.1. Study area and Malaria incidences

The study area is located in the south western part of the central highlands region of Viet Nam, geographically defined between 11°45’ to 12°50’ northern latitudes and between 107°13’to 108°10’ eastern longitudes (Figure 1). The province is characterized by moderate temperature and complex topography that spatially varies from 600m to 1982m. According to provincial information portal (daknong.gov.vn), the province is home for several ethnic minority groups, of which 65% of total population is Kinh (largest community in Viet Nam). The combination of population and physical environment has shaped the livelihoods of local community, education levels as well as attitudes towards disease control and prevention.

The prediction of malaria susceptibility is mostly influenced by input databases. The proper selection of input data affects prediction accuracy how malaria incidences spatially vary. In fact, there are two way to measure malaria occurrences, in which malaria occurrences are measured by point-based locations as in [2, 3] or aggregated data (polygon – based aggregated data) as in [4]. The first manner requires exact coordinates of individual surveys and prediction map are usually measured for every single locations. The second one use average data within certain boundaries (administrative boundaries are usually used) and risk probability is unique for the whole polygon.

Due to limitation in data collection relating to malaria prevalence in the study area, we used point data representing malaria incidences during 2016 and the first two months of 2017. Weekly reports were gathered at Dak Nong preventive Medicine Center, Daknong department of health, in which 62.784 persons had been tested and 125 were diagnosed to be positive with P.f, 118 cases were positive with P.v. Cases with locational information, such as house addresses were geo-referenced basing on their relative positions to road network. The

other cases with limited positional information, additional survey was carried out to provide geographical references.

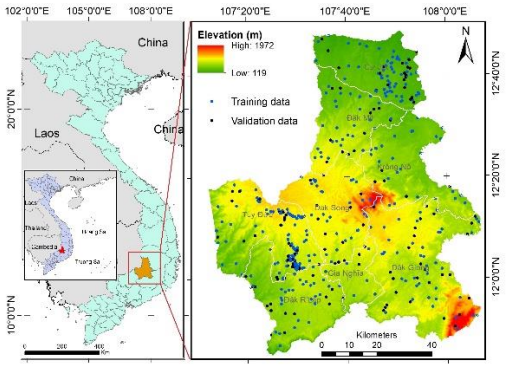

Figure 1. Study area.

Since the model produces binary classes that measure exposure probability to malaria transmission, it is required to have collection of non-infected points. We presumed that the probability decreased as distances from human settlement area increased, so that the same number of presumed non-infected points were randomly selected from the study area. Non-residential area was used as constrained boundary. Totally, overall distribution of 486 points were selected and plotted upon elevation layers as showed in (Figure 1).

2.2. Controlling factors

Since malaria is transmitted by mosquito, it is scrutinized to be sensitive to variations in environmental and socio-economic conditions with regard to living condition of mosquitos and burden for disease prevention activities. Elevation-derived data, vegetation cover, location of water bodies, climatic factors are usual parameters in community disease researches. On the other hand, socio-economic group reflects livelihood condition of local communities and community adaptability to cope with disease transmission risk.

Decision to select appropriate variables for malaria modeling is crucial step to ensure predictive capability of final models. Through screening the literature, we came up with thirteen variables that can be grouped into two groups. The first physical environment group consists of topographic elements namely Digital Elevation Model (DEM), Slope, Aspect and climatic factors such as Rainfall, Temperature, and Humidity. In fact the spatial variation of malaria is highly dependent on climatic factors, in which the transmission varies depending on seasons, rainfall magnitudes, temperature fluctuation, particularly under impact of climate change. The study area is characterized by two distinguished season: dry season from December to May and rain season from June to November. This conditions have impact to vegetation cover and surface temperature and consequently influences how mosquito grows. Currently, this data is extractible from remotely sensed data. In this study, Land Surface Reflectance products of Landsat 8 OLI scene captured in March, 2017 was downloaded from www.earthexplorer.usgs.gov. Several derivable index images from this Landsat that can be used to measure vegetation cover, are Normalized Difference Vegetation Index (NDVI), Normalized Difference Moisture Index (NDMI), Normalized Difference Built-up Index (NDBI). We measured correlation values between each pairs of all three index images and found that there were high correlation between NDVI/NDMI and NDVI/NDBI. Therefore we choose to keep NDVI as it is considered as the most popular index to study vegetation.

In addition to average temperature, Land surface temperature was also measured from the same Landsat dataset. It was converted to Top of Atmospheric spectral radiance, and then to At-satellite brightness temperature at Kevin scale and finally to surface temperature.

The second group of controlling factors demonstrates relationship between human and physical environment that had been studied by [4]. The selection of these factors depends on scale of malaria research in term of point-based study or polygon-based study. Since we focused on the occurrences of malaria, administrative-based aggregated data such as population density, number of raised animals…were not

suitable to be assigned to single locations. Instead, we measured distances to certain types of landuse/landcover with a presumption that the probability of being infected decreases if the distances to those landuse types increase or by versus. Four type of land uses were extracted from 2015-Landuse map namely Residential Land, River, Forest, Wetland, and Locations of Hospital and euclidean distances were calculated.

Using DEM as base raster reference, all thirteen variables were converted into similar data structure at 30x30m resolution in WGS1984, UTM zone 48 projection. All variables are showed in (Figure 2).

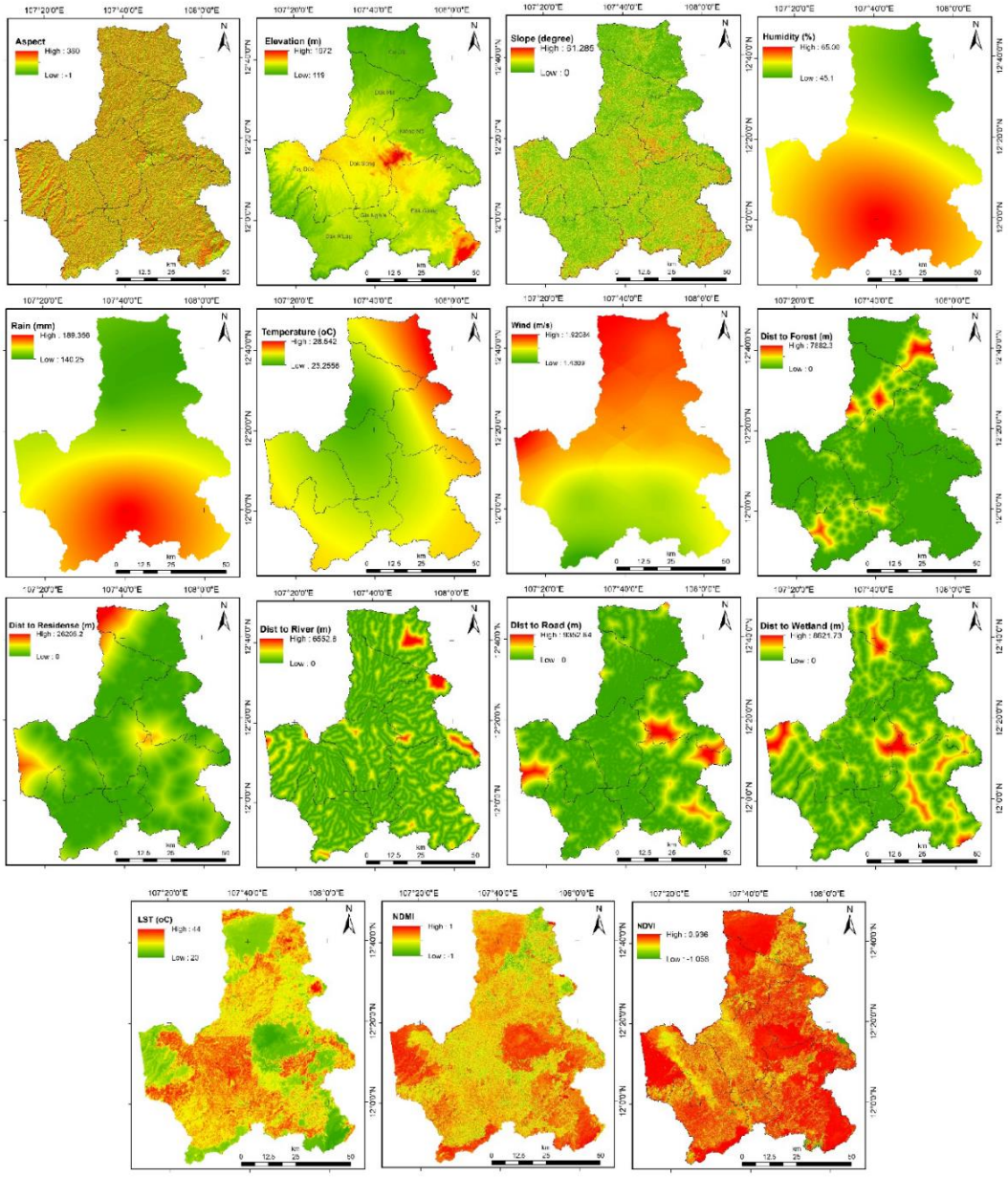

Figure 2. Controlling factors.

2.3. Methods

Since application of data mining techniques in malaria susceptibility mapping is still rare, particularly hybrid method that combines single classifier and an optimization algorithm. This study verifies the capability of simulated annealing optimization in selecting the optimal parameters for Anfis through minimizing the Root Mean Square Error as the objective functions.

Adaptive Fuzzy Inference System (Anfis)

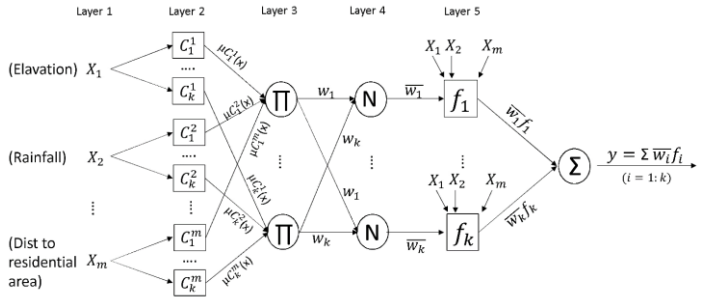

Figure 3. Adaptive Neuro-Inference System

This techniques was first introduced in early 1990s and has been widely used in variation of research topics. Anfis takes advantages of neural network and Takagi-Sugeno/ Mandanni rules in fuzzy logics.

Simulated Annealing

Taking idea of the state of physical process of crystallization aiming at bring the state to minimum energy state, SA was developed to minimize or maximize the global optimum of a function [5]. The optimization process involves permutation of new position that inspires new state with new energy value. This new value is compared to the previous one by pre-defined conditions. If passed, the new state is kept as current state and the iteration continues until meeting maximum number of iteration or desirable energy value. Typical pseudocode presents simulated annealing heuristic as follow:

- Start initial state with value = f_0
- i = 1
- Repeat until Lmax iteration or State level reached
  Pick a random state
  If $f_i < f_{i-1}$ then value = f_iElse
  If exp $\left(\frac{f_{i-1}-f_i}{s_{i-1}}\right)$ > random[0,1] then value = f_i
  $s_i = r * s_{i-1}$
  i = i + 1
- Ouptut: the final state with valuef_i

3. Proposed S-Anfis for malaria susceptibility mapping

3.1. Dataset standardization

Depending on characteristics of data mining algorithms, real values of input datasets might be directly used as in [6] or can be classified into classes as in [7] before further analysis. Normally, for the first choice, variables are measured in different units and scales. It is difficult to use this type in some classifiers or performance of classification model might be reduced. Decision to choose the second type depends on how many classes are determined and how to select threshold values to separate the classes. To some extent, this type generalizes nature of dataset and data detail might be lost. In this study, we used absolute value for the dataset and standardize it into similar unit by using this conversion equation.

$$x_{istandardized} = (x_i - \min)/(\max - \min)$$

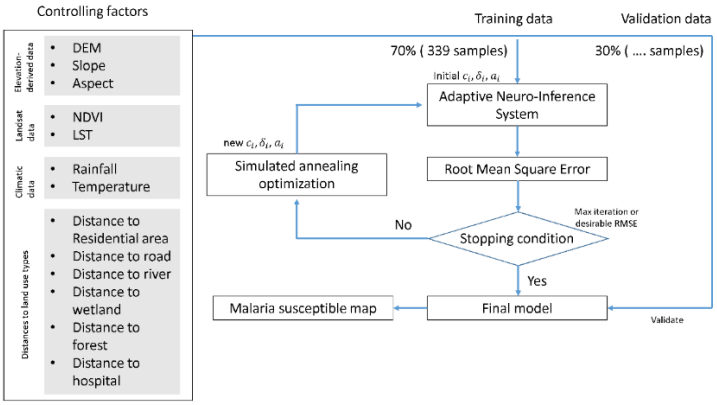

Figure 4. Simulated annealing diagram.

3.2. Initialization of S-anfis

Proposed workflow of S-Anfis is showed in (Figure 4), in which 448 samples were divided into two packs: 70% for training data and 30% for validation. Each sample consisted of 13 controlling factors that were clearly defined in above section (Figure 2). One of the key issues for good performance of S-Anfis is a proper selection of number of rules (or numbers of clusters prior to further processes). Normally, a clustering algorithm is used to define number of clusters if there is no prior understanding of the dataset. This algorithm usually generates high number of clusters that makes model complicated and time-consuming. Literature has showed that by reducing the clusters, model performance will be increased [7]. Through several trials by comparing RMSEs we came up to alternatively run the model with 4,5,8 clusters. The best performance would be selected to produce malaria susceptible map.

One of the options in running the model is to define constraint bounds for parameters. Since value ranges of all variables are limited within [0,1]. As a consequence, a_i, b_i, c_i are also fallen within the similar [0,1] range. Parametersp_i of linear transformation in layer 5 have no bounds, but we decided to limit those within [0,1] for easy calculation.

On the other hand, the Simulated annealing required proper selection of initial parameters, in which initial temperature, temperature cooling function are the most important parameters. These values define acceptance probability of new states. Higher initial temperature avoids sudden jump of accepted new state. Through several trial, we finally used default value for initial temperature at 100, exponential function for temperature cooling process and maximum iteration at 300. The model started with initializing a_i, b_i, c_i, p_i and those parameters were used to generate RMSE for the first iteration. The result was checked if it met predefined threshold or number of iteration exceeded 300. The model continued until stopping condition was met and the final model was validated by validation data.

(Figure 5) shows decreasing trend of RMSE values since the best function values of RMSE were plotted again each iteration. RMSEs had sudden jumps in all three tests and kept unchanged after around the 200th and the 250th iteration. Models with 5 clusters resulted in smallest RMSE values and were used for generating malaria susceptible map (Figure 7).

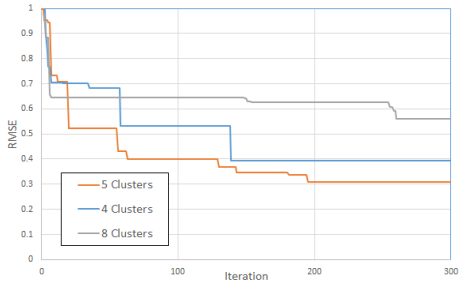

Figure 5. RMSE after 300 iterations.

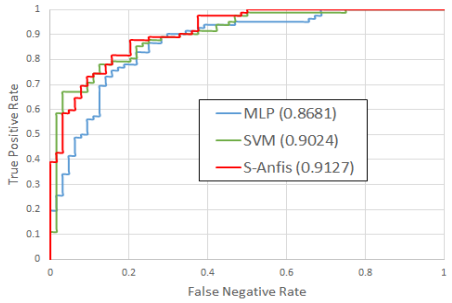

Figure 6. ROCs and AUC values for validation data.

3.3. Performance assessment

For accuracy assessment, Receiver Operating Curve (ROC), Area under ROC (AUC), Cost Curve are widely used for performance assessment of classifications models. (Figure 6) shows ROC curves by validation data for S-Anfis and two benchmark classifiers Support Vector Machine (SVM) and Multilayer Perceptron network (MLP). The results shows that the proposed model out-performed both SVM and MLP in all indications as showed in (Table 1). RMSE rapidly decreased in the first 120 iterations and kept horizontal trend from that point with stable value at 0.265. This value was lower than two RMSEs of two benchmark SVM and MLP.

Table 1. Performance comparison by validation data

Statistical indicators	**MLP**	**SVM**	**S-Anfis**
Kappa statistic	0.541	0.621	0.653
Mean absolute error (MAE)	0.236	0.273	0.239
Root mean squared error (RMSE)	0.430	0.364	0.335
Relative absolute error (%)	47.04	54.36	47.64
AUC	0.868	0.902	0.912

4. Discussions and remarks

The selection of proper variables significantly contributed to the performance of the proposed model. In fact, in many researches focusing on spatial variations of malaria, social – economic factors were have been scored with highest predictive capabilities among other. Normally, those variables were used as aggregated data that provided average value across administrative boundary. This summation, however, results in inaccurate variation patterns as every location within predefined boundary has the same probability values. This study used individual locations of malaria cases to produce susceptible maps providing probability of each pixel within study area. Thirteen variables were selected, of which distances from man-made features can be classified as social – economic factors. Population data (including demography, density) was valuable information but was not put into input database, because there was no significant way to assign those values into single locations. Instead, distance to roads could be used as replacement to population density as the local communities (as well as the Vietnamese) tend to live as close to the roads as possible.

Simulated annealing is single solution - based solution for searching for global optimal, in which model performance is improved over the course of iterations. The main goal of this paper was to investigate whether the combination of Anfis and simulated annealing was capable for optimizing large number of parameters and for solving non-linear functions. Since the objective function (RMSE in this case) consists of premise and consequence parameters that vary depending on number of clusters defined in initial stages. With 5 clusters and 200 parameters, the objective function was successfully solved.

For the second verification in optimizing non-linear optimization problems, two benchmark classifiers MLP and SVM were selected and run with the same training and validation dataset. The two classifiers are widely used in non-linear problems [8]. The goodness-of-fit of two classifiers are dominated by model complexity, such as number of hidden layers in MLP or Kernel function parameters in SVM. By using Grid search techniques, two classifiers with optimal parameters were trained and validated with similar training and validation datasets. Performance comparison of S-anfis model with two benchmark classifiers by using

Kappa index, RMSE, ROC curve indicated that S-anfis outperformed the two in all indicators (Figure 6), (Table 1)

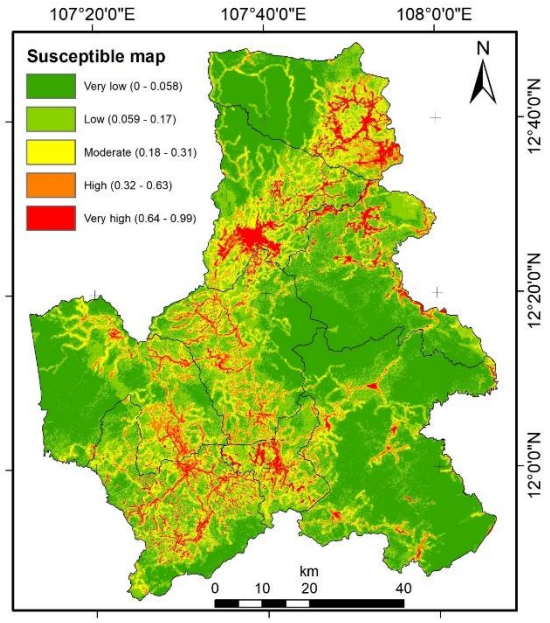

Figure 7. Susceptible map by S-Anfis.

Technically, the selection of Simulated annealing parameters, for instance, initial temperature, temperature decreasing function, function to generate new points only impact how the shape of the plot and how fast this model converge. From several trials, we found that even with different parameters, the models came to similar value after certain iterations. Another aspect should be taken into consideration is how membership function in Anfis is defined. In this study, Bell-shaped function was selected among two other, including Gauss and Sigmoil distribution.

5. Conclusion

This paper filled up the literature in spatial modelling of epidemiology studies, in which a classifier is combined with optimization algorithm. The result of the hybrid model shows a significant improvement of this combination against two benchmark classifiers in all comparing criteria such as RMSE, Kappa, Mean Absolute Error, AUC. Since only Simulated annealing was used in the study, performance of model might be improved if other optimization algorithms are employed such as population-based optimization. More research on this direction should be performed in the future.

The hotspot modelling based on hybrid model shows important risk factors relating to variation of socio-economic and environment condition. The output map provides preliminary understanding of susceptibility levels of the disease in the study area and it can be used as one of important indicators in malaria control and elimination program.

Acknowledgments

This research is funded by the Viet Nam National University, Hanoi (VNU) under project number QG.17.20

References

[1] WHO, World Malaria Report 2016, Geneva, 2016.

[2] M. M. Ndiath et al., "Application of geographically-weighted regression analysis to assess risk factors for malaria hotspots in Keur Soce health and demographic surveillance site," Malaria Journal, vol. 14, pp. 463, 11/18

[3] Q.-T. Bui et al., "Understanding spatial variations of malaria in Vietnam using remotely sensed data integrated into GIS and machine learning classifiers," Geocarto International, pp. 1-15, 2018.

[4] Y. Ge et al., "Geographically weighted regression-based determinants of malaria incidences in northern China," Transactions in GIS, pp. n/a-n/a, 2016.

[5] N. Metropolis et al., "Equation of State Calculations by Fast Computing Machines," The Journal of Chemical Physics, vol. 21, no. 6, pp. 1087-1092, 1953/06/01, 1953.

[6] N. Mathur, I. Glesk, and A. Buis, "Comparison of adaptive neuro-fuzzy inference system (ANFIS) and Gaussian processes for machine learning (GPML) algorithms for the prediction of skin temperature in lower limb prostheses," Medical Engineering & Physics, vol. 38, no. 10, pp. 1083-1089, 2016/10/01/, 2016.

[7] D. Tien Bui et al., "A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area," Agricultural and Forest Meteorology, vol. 233, pp. 32-44, 2017/02/15/, 2017.

[8] D. Tien Bui et al., "GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks," Environmental Earth Sciences, vol. 75, no. 14, pp. 1-22, 2016.

Tích hợp hệ thống suy luận mờ (Anfis) và thuật toán tối ưu hóa Simulated annealing trong nghiên cứu nguy cơ sốt rét tại tỉnh Đắk Nông, Việt Nam

Bùi Quang Thành

Khoa Địa lý, Trường Đại học Khoa học Tự nhiên, ĐHQGHN, 334 Nguyễn Trãi, Hà Nội, Việt Nam

Tóm tắt: Adaptive Neuro-Inference system (Anfis) được sử dụng nhiều trong các ứng dụng phân loại nhị phân. Phương pháp này thường xuyên được sử dụng cùng với thuật toán tối ưu hóa nhằm xác định các tham số tối ưu cho Anfis. Nghiên cứu này thử nghiệm thuật toán Simulated Annealing (SA) và Anfis trong nghiên cứu nguy cơ sốt rét tại tỉnh Đắk Nông, Việt Nam. Để đánh giá độ chính xác của mô hình, thông số ROC được sử dụng cùng với một số chỉ số thống kê khác. Kết quả nghiên cứu cho thấy độ chính xác của mô hình đề xuất so với các mô hình dùng để so sánh như sau S-Anfis (AUC = 0.912, RMSE =0.335) Support Vector Machine (AUC = 0.902, RMSE =0.364), Multiple LayerPerceptron (AUC = 0.868, RMSE =0.430). Kết quả này cho thấy mô hình kết hợp giữa SA và Anfis cho kết quả tốt hơn các phương pháp khác, và có thể được sử dụng cho nghiên cứu nguy cơ sốt rét tại các địa phương khác tại Việt Nam

Từ khóa: Anfis, Simulated annealing, Sốt rét.