TOWARDS CLASSIFICATION OF SYMPTOMATIC AND CLIMATIC BASED MALARIA PARASITE-COUNT

R.G. Jimoh[1], *O.A. Abisoye,[2] M.M.B. Uthman[3]

[1] Department of Computer Science, Faculty of Communication and Information Science,
University of Ilorin, Nigeria.
jimoh_rasheed@unilorin.edu.ng, jimoh_rasheed@yahoo.com
[2] *Department of Computer Science, School of Information and Communication Technology,
Federal University of Technology, Minna, Niger State, Nigeria.
o.abisoye@futminna.edu.ng, opeglo@yahoo.com.au
[3]Department of Epidemiology and Community Health, Faculty of Clinical Sciences, College of Health Sciences,
University of Ilorin, Nigeria.
uthman.mb@unilorin.edu.ng, uthmanmb@yahoo.com

Abstract

Good health is one of the most important things in life. Several diseases affect the proper functioning of human's health; one of such common disease is malaria. Malaria is a leading public health problem in the developing countries and Nigeria, leading to high morbidity and mortality and huge cost for diagnosis, treatment and control. The insurgency of malaria diseases has pushed the need to develop computational approaches for predicting the severity of malaria diseases based on symptoms and climatic factors. The prediction of the occurrence of malaria disease and its outbreak will be helpful to take appropriate precaution measures. The existing predicting models examine binary cases of malaria, prone to error, and suffer from overfitting due to large number of parameters to fix. This paper proposes a Support Vector Machine (SVM) with best activation function to determine the rate of malaria transmission. This paper aims to study the components of learning parameters in multiclass Support Vector Machine (SVM), study optimal separation hyperplane, review SVM classification and generate SVM malaria model. *Monthly averages of rainfall, temperature, relative humidity and malaria serves as the input variables. Apart from classification, the future work will be based on using SVM for other machine learining technique functions such as pattern recognition, regression analysis and feature selection.*

Keywords: Malaria, Parasite-Count, Support Vector Machine (SVM), Classification, Prediction, Symptomatic

I. Introduction

Infections are diseases caused by virus, bacterial and fungal organisms. Such diseases could be sexually transmitted, air borne or water borne, examples are glaucoma, tuberculosis, measles, chickenpox, influenza, malaria e.t.c. The World Health Organization has enrolled several programmes to reduce malaria infection but we still suffer from its insurgency [1]. Hundreds of millions of dollars has been donated by Bill and Melinda Gates Foundation funded by Bill Gate to reduce malaria infection. Since 2003, Funds devoted to malaria reduction has been doubled. Emerging efforts and ideas is to disable the disease by combining virtually every known malaria-fighting technique, from the ancient (Chinese herbal medicines) to the old (bed nets) to the ultramodern (multidrug cocktails). At the same time, malaria researchers are pursuing a long-sought, elusive goal: a vaccine that would prevent the transmission of malaria to man despite exposure to mosquitoes [2].

Malaria transmission is affected by environmental and climatic factors; warm humid and rainfall favour the transmission of malaria. Several non-climatic factors, such as human/behavioural factors can also affect the pattern of malaria transmission and the severity [3].

A high prevalence of malaria cases is seen every year and there are difficulties in predicting its future occurrence and analysis of its possible threats. So, the need for a machine learning method arises. Machine learning processes the data and automatically finds structures in the data, i.e. learns. The knowledge about the extracted structure can be used to solve the problem at hand.

.

II. Literature Review

Conventional microscopy used in malaria diagnosis occasionally proved to be time consuming and sometimes results are difficult to reproduce. Alternative diagnostic techniques which yield superior results are quite expensive.

A fully automatic system for detection of infected erythrocytes from blood images and parasite life stage identification counting was developed. But in the study the classifications of malaria parasite on thin blood Imaging system is time consuming. So, an Otsu Threshold, SVM Binary Classifier and SVM multi classifier was developed for the counting. SVM binary classifier gives 96.26% sensitivity and 99.09% specificity results on 71 images. SVM multiclass classifier (i.e. RBF kernel) gives 96.42% accuracy for correct identification rate of life stage of parasite [4].

A study focuses on morphological characteristics and novel threshold technique. On 70 images of blood cells ANN gave accuracy of 78.53% and running time of 6.58 seconds while SVM gave accuracy of 98.25% [5]. Support Vector Machine (SVM) and Naïve Bayes classifier and two feature extraction techniques Discrete Wavelength Transform (DWT) and Gray-Level Co-Occurrence Matrix (GLCM) feature extraction technique were developed. The study revealed that Artificial Neural Network (ANN) classifiers are suitable for detection and classification of Plasmodium parasites into their respective stages and species [6].

In the SVM utility with adaptic thresholding to detect the malarial cases for other geographical regions was not considered. Also, comparisons of the SVM with the other new methods such as extreme learning machines were not considered. Several literatures indicate that the proposed SVM model provides more accurate forecasts compared to the other traditional techniques [7].

Modeling goal is to choose a model from the hypothesis space, which is closest to the underlying function in the target space [9]. The outcome of learning process is being affected by the choice of an error measure. This paper aims to study the components of learning parameters in Support Vector Machine (SVM), review SVM classification, study optimal separation hyperplane for SVM Model and generate SVM malaria model

III. Methods

The existence of mathematical, statistical and machine learning algorithms simplifies modeling approaches. A machine learning algorithm to generate support vectors is called a Support Vector Machine. It is mainly used in prediction, classification, feature selection, pattern recognition and regression analysis. SVM performance relies on its appropriate parameters selection which is very complex in nature and quite hard to solve by conventional optimization techniques. Severally, SVM has been proved that its applications show superior

performance compared to prior developed methodologies such as neural network and other conventional statistical applications [10-13].

SVM applications span through variety of fields such as computing, hydrology, medicine and environmental researches [14 -15].

a. Components of SVM Learning

Main Components of SVM Learning are:

i. *X:* Instance space or input space where an observation x belongs to this space

ii. *Y:* Output Space or label space which is the set of all output where an output y belongs to this space. Examples: Binary Classification y = {±1}, Multiclass Classification y = {1…K}, Regression y∈R

iii. l: y y= R*: Loss Function (measures prediction error)

iv. *f: X→Y :* Target Function/Model/Hypothesis Class (Set of functions from input space to output space). The unknown target function is the correct formula that correctly mapped x to the corresponding output *y*.

v. *Training Data SetD(**D**):* This is the set of historical records ***x*** with their corresponding output ***y***: *{(x_1,y_1),…,(x_n,y_n) where $y_n = f(x_n)$,* n= 1,2,…,*l*.

vi. *Learning Algorithm (g:* A learning algorithm uses the data set *D* to pick a formula *g: X→Y* that approximates *f*. the algorithm chooses *g* from a set of hypothesis set *H*(candidates formula under consideration)

vii. *g(x):*the hypothesis that the learning algorithm produced is represented mathematically as *g(x) = sign ((w_i,x)+b*) where w is the weight and b is the bias. Fig 1. below portrays the components of learning problem.

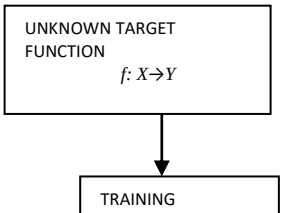

Fig 1: Basic set up Learning Problem [8]

b. Support Vector Machine Classification

Support Vector Machine classifies binary class problems without loss of generality but also classifies multiclass problems by adopting a unique algorithm. Its primary goal is to separate the two classes by a function induced from an available example by a classifier called '***optimal separation hyperplane'.*** This optimal classifier separates the data by finding the maximum margin among the two classes of data.

An SVM training algorithm designs a model that predicts whether a new data mapped and fitted in a category or the other or learned from historical examples. SVM models take a set of input data and predict for each instance the possible output which makes them non probabilistic binary linear class [16]. The SVM algorithm operation is based on finding the hyper-plane with the highest minimum distance to the training examples. Thus, the optimal separating hyperplane maximizes the margin of the training dataset.

IV. An Optimal Separation Hyperplane for Malaria

In the problem of classifying this set of training vectors of two classes of malaria patterns

$M = \{(x_1,y_1),\ldots,(x_n,y_n)\}$, $x_i \in R_n$ $y_i \in \{-1,1\}$ (1)

Where $\{x_i,y_i\}_1^n$ is the assumed training data sets where $x_i \in R_n$ is the input space vector of real number sample data and $y_i \in \{-1,1\}$ is the target value.

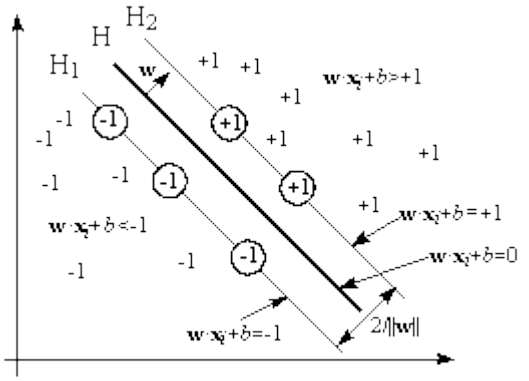

Fig 3: **An optimal separation hyperplane for malaria**

A linear classifier that can perfectly separate them as shown in Fig 3 above is deduced below as

$$(w,x)+b = 0 \quad (2)$$

This linear classifier function is

$$f(x) = \sum_{i=1}^{n} w_i, x_i + b \quad (3)$$

or

$$f(x) = w\phi(x) + b \quad (4)$$

From (3), $\phi(x)$ denotes high dimensional feature space that mapped the input space vector x, w is a normal vector, b_i is a scalar quantity or bias

A linear classifier that can optimally separates the malaria data is the one that minimizes

$$M(w) = \frac{1}{2}||w||^2 \quad (4)$$

Equation 4 is independent of b provided this constraint in Equation 5 below is satisfied.

$$y_i[(w,x_i)] + b \geq 1, \quad i=1,\ldots,l \quad (5)$$

The optimization problem of equation 4 solution above subjected to the constraint in equation 5 is given by the introduction of Langrange function

$$M(w,b,\alpha) = \tfrac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (y_i [(w, x_i) + b] - 1), \quad (6)$$

where α is the Langrange multiplier. The Langragian will be minimized with respect to *w,b* and maximized with respect to $\alpha \geq 0$. Thus equation 6 above is transformed to its dual problem as thus

$$max_\alpha Mal(\alpha) = max_\alpha [(\min M(w,b,\alpha)] \quad (7)$$

The minimum with respect to w and b of the Langrangian, M, is given by,

$$\frac{\partial M}{\partial b} = 0 \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (8)$$

$$\frac{\partial M}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad (9)$$

Thus, the dual problem from equation (7),(8) and (9) above is given by

max_{α}Mal(α) = max_{α} $-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j\,y_i\,y_j\,(x_i,x_j)$ + $\sum_{k=1}^{n}\alpha_k$, (10)

Hence, the solution to the problem is given by

α^{mal} = arg min_{α} $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j\,y_i\,y_j\,(x_i,x_j)$ - $\sum_{k=1}^{n}\alpha_k$, (11)

with constraints

$\alpha_i \geq 0$ i = 1,…,n $\sum_{j=1}^{n}\alpha_j\,y_j = 0$ (12)

Thus Equation 11 with constraints in equation 12 determines the Lagrange Multipliers and the optimal separating hyperplane is given by

$$w^{mal} = \sum_{j=1}^{n}\alpha_i\,y_i\,x_i \quad (13)$$

$$b^{mal} = \frac{1}{2}(w^{mal}, x_c + x_s) \quad (14)$$

Where $x_c + x_s$ are any support vector from each class satisfying $\alpha_c, \alpha_s > 0$, y_c = -1, y_s = 1.

To generalize the optimal separating hyperplane method we introduced positive variables $\varepsilon_i \geq 0$ and a penalty function

$$F_{\sigma}(\varepsilon) = \sum_i \varepsilon_i^{\sigma} \quad \sigma > 0 \quad (15)$$

Where the $\varepsilon_{,i}$ are the measure of misclassification errors.

From equation 1.4 thus we have

$$M(w, \varepsilon) = \frac{1}{2}||w||2 + C\sum_i \varepsilon_i^{\sigma} \quad (16)$$

Subjected to the constraint in 1.5

$y_i[(w,x_i)] + b \geq 1 + \xi_i, \quad \varepsilon_i \geq 0,\ i=1,\ldots,l$ (17)

Where C is a regularization constant that has to be determined to prevent overfitting or reflect the knowledge of the noise on the data and minimized the training error.

By applying Langrangian theory, the Lagrangian will be minimized with respect to w,b, ξ and maximized with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as done in equation (15-17) above. Thus the dual problem is

max_{α}Mal(α) =max_{α}[(min M(w,b,α,ξ, $\boldsymbol{\beta}$)] (18)

The minimum with respect to *w,b,* ξ and of the Langrangian, M, is given by,

$$\frac{\partial M}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i\,y_i = 0 \quad (19)$$

$$\frac{\partial M}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n}\alpha_i\,y_i\,x_i \quad (20)$$

$$\frac{\partial M}{\partial \xi} = 0 \Rightarrow = \alpha_i\,\beta_i = C \quad (21)$$

Thus, the dual problem from equation (19) (20) and (21) above is given by

max_{α}Mal(α) = max_{α} - $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j\,y_i\,y_j\,(x_i,x_j)$ + $\sum_{k=1}^{n}\alpha_k$, (22)

Hence, the solution to the problem is given by

α^{mal} = arg min_{α} $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j\,y_i\,y_j\,(x_i,x_j)$ - $\sum_{k=1}^{n}\alpha_k$,

(23)

with constraints

$0 \leq \alpha_i \leq C \quad i = 1,\ldots,n \quad \sum_{j=1}^{n} \alpha_j y_j = 0$ (24)

To make generalization into high dimensional feature space training vectors x_i is mapped into a higher dimensional space by a kernel function *K*. Then SVM finds a linear separating hyper plane with the maximal margin in this higher dimensional space. So $k(x_i, x_j) = \phi(x_i)\,\phi(x_j)$ is called the kernel functions. There are number of kernels that can be used in SVM models. These include linear polynomial, RBF and sigmoid [17]: $\phi = (x_i \times x_j)$

Linear, $(\gamma x_i x_j - coeff)^d$ Polynomial,

$\exp(-\gamma |x_i - x_j|^2)$ RBF

$\tanh(\gamma x_i x_j - coeff)$ Sigmoid

Thus far, the optimization problem in equation (23) now becomes

$\alpha^{mal} = \arg\ min_{\alpha}\ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j\ K(x_i, x_j) - \sum_{k=1}^{n} \alpha_k$,

(25)

with constraints

$0 \leq \alpha_i \leq C \quad i = 1,\ldots,n \quad \sum_{j=1}^{n} \alpha_j y_j = 0$

(26)

Where $K(x_i, x_j)$ is the kernel function mapping(x_i, x_j) into feature Space *X*. Where $Ks(x_i, x_j)$ is an explicit dot product of **φ :** $K(x_i, x_j) =$ **[(φx$_i$), (φx$_j$)].**

V. SVM Malaria Model

In SVM the malaria data to be classified can be formally written as:

$M = \{(x_1,y_1,z_1), \ldots, (x_n,y_n,z_n)\}$, $x_i \in R_n$, $y_i \in \{-1,1\}$, $z_i \in Z$

(27)

Where *Z* is a set of malaria membership? Recall that ξ_i is the measure of classification error, So in **SVM** for malaria Z_i ξ_i is the new classification error, so we have to minimize this optimal problem:

$M(w, \varepsilon) = \frac{1}{2}||w||2 + C\sum_i \varepsilon_i^{\sigma}\ z_i$ (28)

Subjected to the constraint in 5

$y_i[(w,x_i)] + b \geq 1 + \xi_i, \quad \varepsilon_i \geq 0,\ i = 1,\ldots,l$ (29)

By applying Langrangian theory, the Lagrangian will be minimized with respect to w, b, ξ and maximized with respect to α and β as done in equation (18 -21) above. Thus the dual problem is

$max_{\alpha} \text{Mal}(\alpha) = max_{\alpha}[(\min M(w,b,\alpha,\xi, \beta)]$ (30)

The minimum with respect to *w,b,* ξ and of the Langrangian, M, is given by,

$\frac{\partial M}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i$ (31)

$\frac{\partial M}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i z_i$ (32)

$$\frac{\partial M}{\partial \xi_i} = 0 \Rightarrow = \alpha_i \beta_i = z_i \text{ C- } \alpha_{i-} \beta_i \quad (33)$$

Thus, the dual problem from equation (31),(32) and (33) above is thus given by

$$max_{\alpha} \text{Mal}(\alpha) = max_{\alpha} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i, x_j) + \sum_{k=1}^{n} \alpha_k , \quad (34)$$

Hence, the solution to the problem is given by

$$\alpha^{mal} = \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \, \text{k}(x_i, x_j) \quad (35)$$

$$\alpha^{mal} = -\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \, \text{K}(x_i, x_j) \quad (36)$$

with constraints

$$0 \leq \alpha_i \leq z_i C \quad \text{i} = 1,\ldots,\text{n} \quad \sum_{j=1}^{n} \alpha_j y_j = 0$$

From equation (13) and (14) above the bias $\boldsymbol{b^{mal}}$ is computed by the use of two support vectors $\text{x}_c + \text{x}_s$ thus we have

$$\boldsymbol{b^{mal}} = -\frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i [\text{K}(x_c, x_s) \, \text{K}(x_c, x_s)] \quad (37)$$

5.0 Conclusion

Prediction of the transmission and severity occurrence of malaria and parasite count to capture its future occurrence, effectively plan control and reduce its burden is very vital. Support Vector machine (SVM) model gives the nearest response to prediction call because it has the potential of combining human heuristics into computer assi sted decision. In this paper, we have proposed an intelligent system Support vector Machine, studied the components of learning parameters in SVM, study optimal separation hyperplane thay gives optimal results and generate SVM malaria model review SVM Classification.

SVM with appropriate activation function will provide a simple way to get a definite conclusion from ambiguous medical data. The future work will be based on using SVM for other prediction characteristics like feature selection. Contributions and Suggestions are welcome at this stage of the research.

References

[1] World Health Organization, Fact Sheet: World Malaria Report, 2015.

[2] National Geographic Society, In the Malaria report titled "Malaria—Bedlam in the Blood", , 2015.

[3] Depinay, C.M. Mbogo., G. Killeen, B. Knols, J. Beier, J. Carlson, J. Dushoff, P. Billingsley, H. Mwambi, J. Githure, et al.. A simulation model of african anopheles ecology and population dynamics for the analysis of malaria transmission, Malaria J. 3 (1) (29), 2004.

[4] S.S. Savkare, & S.P. Narote , Automatic system for classification of erythrocytes infected with malaria and identification of parasite's life stage. *Procedia Technology*, *6*, 405-410, 2012.

[5] S. Annaldas, S.S. Shirgan & V.R. Marathe, Automatic identification of malaria parasites using image processing. *Int. J. Emerging Eng. Res. Technol*, *2*, 107-112, 2014.

[6] T. Chaudhari, , & D. Agrawal. Automatic Detection of Malaria Parasites for Estimating Parasitemia: Review Paper, 2015.

[7] S. Ch, S.K. Sohani, D. Kumar, A. Malik, B.R. Chahar., A.K. Nema,... &R. Dhiman,. CA support vector machine-firefly algorithm based forecasting model to determine malaria transmission. *Neurocomputing*, *129*, 279-288, 2014.

[8] Y.S. Abu-Mostafa, M. Magdon-Ismail, &H.T. Lin, *Learning from data*. AMLBook. (2012).

[9] S.R. Gunn, Support vector machines for classification and regression.*ISIS technical report*, *14*, 1998.

[10] AH. Sung , S. Mukkamala, Identifying important features for Intrusion Detection using support vector machines and neural

networks. In: Applications and the Internet, 2003. Proceedings. 2003 Symposium on; 2003. IEEE

[11] V. Vapnik, SE. Golowich, , A. Smola, Support vector method for function approximation, regression estimation, and signal processing. Adv Neural Inform Process Syst :281–7,1996.

[12] Kasra Mohammadi, Shahaboddin Shamshirband, Chong Wen Tong, Muhammad Arif Dalibor Petkovic, Sudheer Ch,. "A new hybrid support vector machine–wavelet transform approach for estimation of horizontal global solar radiation", Energy Conversion and Management. 2014. www.elserivier.com

[13] S.Rajasekaran, S. Gayathri, and TL Lee,. Support vector regression methodology for storm surge predictions. Ocean Eng ;35(16):1578–87, .2008.

[14] T. Asefa , M. Kemblowski, M. McKee and A. Khalil, Multi-time scale stream flow predictions: the support vector machines approach. J Hydrol;318:7–16,2006.

[15] Y. Ji, and S. Sun, 2013. Multi task multi class support vector machines: model and experiments, Pattern Recognition 46(3) 914–924.

[16] S. Ch, n. Anand, B.K. Panigrahi, & S. Mathur, Streamflow forecasting by SVM with quantum behaved particle swarm optimization. *Neurocomputing, 101*, 18-23, 2013.

[17] Y. Rejani & S.T. Selvi ,Early detection of breast cancer using SVM classifier technique. *arXiv preprint arXiv:0912.2314*, 2009.

Profiles

Jimoh Rasheed Gbenga is currently an Acting Dean of Faculty of Communication and Information Science (FCIS), University of Ilorin, Nigeria. He attended Universiti Utara Malaysia, Malaysia where he got Ph.D. in Information Technology. His research interests are: Information Security, Soft Computing and Machine Learning.

Professional Membership:
A member of Computer Professionals[Registration Council of Nigeria]-**CPN**
A member of Nigeria Computer Society of Nigeria (NCS).
A member of IEEE Nigeria Chapter.

Abisoye Opeyemi A. was born in Ogbomoso, Oyo State, Nigeria.. She attended University of Ilorin, Ilorin, Nigeria where she obtained her BSc, Msc, degree in Computer Science,. She is currently a PhD. Student of the same institution. She is major in Computational Intelligence, Machine Learning, Data Mining, and Soft Computing.

She serves as a Lecturer I, in the Department of Computer Science, SICT, Federal University of Technology, Minna, Niger State, Nigeria from May 23rd 2007 Till Date.

Professional Membership: (MPCN) A member of Computer Professionals[Registration Council of Nigeria]-**CPN (30th June, 2010)**

Uthman, Muhammed Mubashir Babatunde is currently a Senior Lecturer, Department of Epidemiology and Community Health, Faculty of Clinical Sciences, College of Health Sciences, University of Ilorin, Nigeria.

He attended University of Ilorin where obtained MB;BS., and MPh degrees. He is a Fellow West African College of Physicians, Faculty of Community Health. His research interests are: Environmental Determinants of Health and Diseases.

Professional Membership:
A member of NMA, MDCAN and APHPN

Funding

This research was supported by TETFund Institutional Based Research Intervention(IBRI), University of Ilorin, Ilorin, Nigeria.