CLASSIFICATION AND FEATURE SELECTION OF SYMPTOMATIC AND CLIMATIC BASED MALARIA PARASITE COUNTS USING SUPPORT VECTOR MACHINE

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

Dynamics of Malaria parasite transmission is complex and been widely studied. Research is needed to find a subset of the original features, that will generates a classifier with the highest possible accuracy. Feature selection improves classifier performance; because some machine learning algorithms are known to degrade in performance when faced with many irrelevant/noisy features. In this paper, Support Vector machine (SVM) with One_against_all algorithm is employed to select optimal features for the multiclass symptomatic and climatic malaria parasite-count. Monthly surveys of malarial incidences cases were collected from sampled health centers in Minna Metropolis, Niger State, Nigeria and served as input variables. Linear, Radial Basis and polynomial kernel function were employed but SVM with radial basis kernel function produced better performance result of 85.60% Accuracy, 84.06% Sensitivity and 86.09% Specificity at optimum threshold value of 0.60. SVM selected optimal features to improve prediction performance and reduces time complexity. The experimental results show the robustness and reliability of the proposed model compared to the previous related models.

Index Term: Malaria, Support Vector Machine (SVM), Feature Selection, Prediction, Symptomatic, Climatic, Multi-class, Parasite-counts

1.0 Introduction

Malaria transmission is site specific due to variations of climatic conditions of a region. Temperature, rainfall, relative humidity variations affects the life cycle of malaria parasite [1]. Other non-climatic factors, such as human/behavioural factors can also affect the spread of malaria transmission and severity [2].

Recent researches focuses on dynamics and complexities of Malaria parasite transmission. Research is ongoing on how the risk of asymptomatic and symptomatic malaria infection changes [3, 4]. Malaria parasite count diagnosis can be asymptomatically or symptomatically low, mild and high. Sometimes, many symptoms of different patient may even

*Corresponding Author: Abisoye Opeyemi A. E.mail: o.abisoye@futminna.edu.ng overlap. A malaria patient cases may even have characteristics of other diseases. Therefore, medical problems cannot be generalized and analyzed by imagination. An Knowledge intensive program should be conducted to integrate this complex network of problems and devise individualized solutions [5]**.**

Consequently, a huge amount of malaria cases which is hard to understand and to interpret by humans are collected every year [6]. So difficulties arises on how to analyse the data and interpret it to reduce or possible eradicate subsequent occurrences. Then, the need for a machine learning (ML) method arises. ML processes the data and automatically learns from the data. The knowledge generated from the extracted infection cases can be used to solve the problem at hand.

Problems being solved by machine learning methods involves classifying observations, predicting values, structuring data (e.g. clustering), compressing data, visualizing data, filtering data, selecting relevant components from data when faced with many irrelevant/noisy features., extracting dependencies between data components, modeling the data generating systems, constructing noise models for the observed data, integrating data from different sensors, using classification and drawing inferences.[7,8]

This paper proposed a Machine Learning (ML) method, Support Vector Machine linear, radial basis, and polynomial kernel function (SVM-rbf) to make control tradeoffs between large datasets, sparsity of data representation and select relevant features from data. This will help to reduce space use when working with a limited amount of system memory.

Feature Selection aim is to select features that leads to a large between class distance and small within class variance in the feature vector space [9]. It finds a subset of the original features, that will generates a classifier with the highest possible accuracy. There are quantitative (continuous), ordinal and categorical (nominal/discrete) types of features. Some classifiers like Naïve Bayes, decision trees, treat categorical and quantitative features differently.

Feature selection gives a better understanding of the data and the classification rule [10, 11]. It avoids computational complexity by reducing the number of features to a sufficient minimum. It also improves classifier performance; because some machine learning algorithms are known to degrade in performance. The theoretical justification to retain the highest weighted features for feature selection was ascertained [12].

2.0 Literature Review

In Sindhwani *et al.,* study, theoretical justification for retaining the highest weighted features has been independently derived in a somewhat different context [12].

Their experiments on text categorization compare the effectiveness of the SVM-based feature selection with that of more traditional feature selection methods. Experimental results indicate that, at the same level of vector sparsity, feature selection based on SVM normals yields better classification performance than odds ratio- or information gain based feature selection when linear SVM classifiers are used [12]. SVM was also used as a classifier that outperforms most of other classification methods on text data [13, 14]. The limitation of the research was the evaluation of their approach on other data sets, perhaps on domains outside text categorization.

Olivier and Sathiya in 2008 evaluated new embedded methods on a number of text classification problems and demonstrate that they are quite superior to a baseline filter method that uses information gain[15]. In parallel works of Obozinsky et al [16] and Argyriou et al [17] a similar model for L1 regularization was developed. They models were applied on multi-task learning and use a block coordinate-wise optimization technique for training.

A research on Support Vector Machine-Firefly Algorithm for malaria diagnosis was conducted in India to classify malaria cases. The motivation was that the performance of SVM mainly depends on its appropriate parameters selection which is very complex in nature and quite hard to solve by conventional optimization techniques. The results indicate that the proposed SVM-FFA model provides more accurate prediction compared to the other traditional techniques. The limitation to the study was that the lead times (such as bi-monthly, quarterly or yearly prediction) were not considered [18].

3.0 Materials and methods

Monthly surveys of malarial incidences were collected from sampled health centers in Minna Metropolis, Niger State. Climatic data consisting of Monthly averages of rainfall, temperature and relative humidity were collected from Nigerian Environmental and Climate Observation Programme(NECOP) Weather Station, Bosso Campus, Federal University of Technology, Minna, Niger state. Each patient has a set of symptoms and MP count known as Patients' malaria data symptoms and lab test results. This Climatic data combined with monthly malaria incidences were considered as input variables was trained and simulated using Microsoft Excel and libSVM in MATLAB 2015a

Sampled hospitals laboratories, Giemsa staining was used for the laboratory tests. The Red blood cells (RBCs), Plasmodium spp, platelets and other artifacts were identified. This Plasmodium spp is measured in count being called Malaria Parasite Count (MPcount).

3.1 One Against all Algorithm

SVM is a binary classifier but the algorithm can be used to solve multiclass problem by introducing One-Against-All Algorithm that captures single handedly each class of the target and compare it with the other classes.

Table 1: One-Against-All

```
Input: Training Malaria Datasets

Output: Optimal Features

1. Begin
2. For counter= 1 to  Size(target,1)
3.     if Target(counter) = 0
           Target(counter) == 0
       Else
           Target(counter) == 1
       End
4. End
```

3.2 Feature Selection Algorithm

The SVM feature selection algorithm was thresholded as shown in Table2 to get the optimum threshold value that will yield best result for the model using the above One-Against-All algorithm.

Table 2: SVM feature Selection Algorithm

```
Input: Training Malaria Datasets

Output: Optimal Features

1.  Begin
2.  Input the Malaria Data Features
3.  Preprocess the data by using the most
    suited normalization method
4.  Divide the data into Training Malaria
    Datasets and Testing Malaria Datasets in
    ratio 70:30
5.  Perform One_Against_All(OAA) algorithm to
    convert Multiclass to Binary class in
    preparation for feeding into SVM
6.  WhileThreshold_Value> =0.100 Step 0.05
    Do
7.  While Accuracy_Instances<= No_of _Runs
8.  Train an SVM
9.  Simulate SVM
10. Recall Simulated SVM
11.  Simulate with Transposed Testing Malarial
    Datasets
```

```
Get Simulation Results
Compute Optimal Features, Accuracy ,Performance   Evaluation
EndWhile
EndWhile
End
```

4.0 Feature Selection

Given a number of features, wrapper method and Support Vector Machine were used to select subset of features that have the greatest predictive power and still carry their class discriminatory properties. The dataset has these prevalent features: Headache (Hd), Fever (F), Dizziness(D), Body Pain(Bp) and Vomiting(V_m). The climatic factor; temperature, relative humidity and rainfall contributing factors to being having malaria are also the combined features This research features is thus restricted to five(5) predominant malarial symptoms and climatic factors

4.0 Results

The Multiclass malaria data was handled by one-against-all algorithm. The result of various classes of SVM Feature Selection with 1200 malaria cases; 840:180:180 were used for Training, Testing and Validation respectively is presented in *Table 3(a), 3(b) and 3(c).* Also the Graph of the Support vectors Vs. Accuracy for SVM_0, SVM_1, SVM_2 are depicted in

Figure 1(a), *Figure 1(b), Figure 1(c),* respectively.

Class 0, Class1 and Class 2 malaria cases were trained, tested and validated with linear, radial basis and polynomial function single handedly. Their results were depicted in *Table 3(a), 3(b) and 3(c).*

Table 3(a): SVM_0 Feature Selection Results

Performance Metrics	SVM_0 (lin)	SVM_0 (rbf)	SVM_0 (pol)
Accuracy (%)	81.67	86.11	86.11
Support Vectors	232 x 8	324x8	208 x 8
True Positive	0	11	15
True Negative	147	144	140
False Positive	0	3	7
False Negative	33	22	18
Sensitivity	0	0.3333	0.4545
Specificity	1	0.9797	0.9524
(FP_R)	0	0.0204	0.0400
(FN_R)	1	0.6667	0.5455
(MSE)	0.7333	0.5556	0.5556
Total Positive	33	33	33
Total Negative	147	147	147
Total	180	180	180

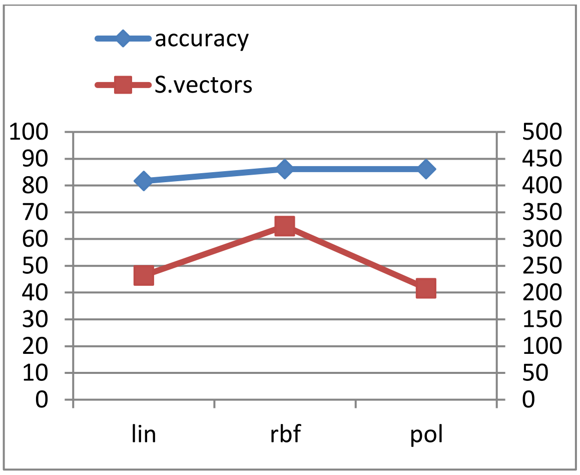

Figure 1(a) Graph of Accuracy Vs Support Vectors for 'SVM_0' Malaria cases

Table 3(b): SVM_1 Feature Selection Results

Performance Metrics	SVM_1 (lin)	SVM_1 (rbf)	SVM_1 (pol)
Accuracy (%)	66.67	80.55	83.89
Supptor Vectors	378 X 8	435 X 8	282 X 8
True Positive	63	60	63
True Negative	57	85	88
False Positive	45	17	15
False Negative	15	18	14
Sensitivity	0.8077	0.7692	0.8077
Specificity	0.5588	0.8333	0.8627
(FPR)	0.4412	0.1667	0.1311
(FNR)	0.1923	0.2308	0.1311
(MSE)	1.3333	0.5556	0.6444
Total Positive	78	78	78
Total Negative	102	102	102
Total	180	180	180

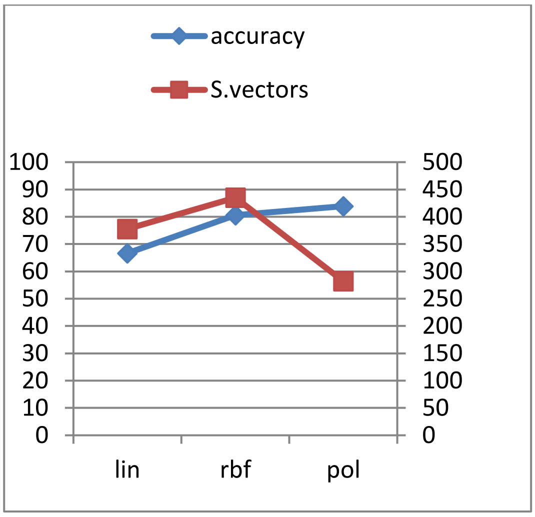

Figure 1(b) Graph of Accuracy Vs Support Vectors for 'SVM_1' Malaria cases

Table 3(c): SVM_2 Feature Selection Results

Performance Metrics	SVM_2 (lin)	SVM_2 (rbf)	SVM_2 (pol)
Accuracy (%)	79.44	85.60	88.89
Support Vectors	195 x 8	308 x8	147x 8
True Positive	52	58	63
True Negative	9	96	97
False Positive	20	15	14
False Negative	17	11	06
Sensitivity	0.7536	0.8406	0.9130
Specificity	0.8198	0.8649	0.8736
(FPR)	0.1802	0.1351	0.1262
(FNR)	0.2464	0.1594	0.0870
(MSE)	0.8222	0.5778	0.444
Total Positive	69	69	69
Total Negative	111	111	111
Total	180	180	180

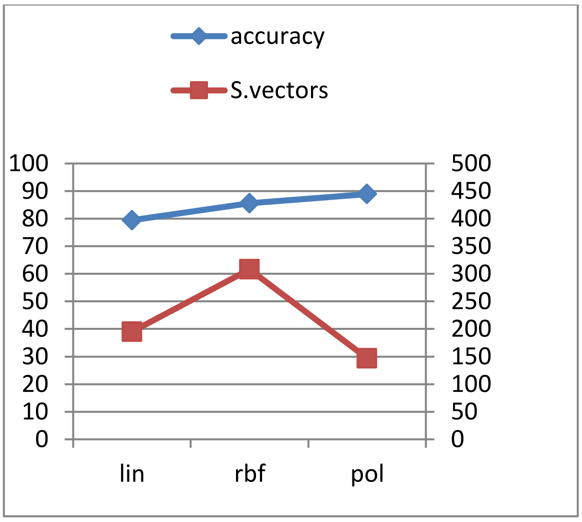

Figure 1(c): Graph of Accuracy Vs Support Vectors for 'SVM_2' Malaria cases

5.0 Discussion

b. **Support Vectors:** Support Vectors are closest data points to the hyperplane. It was observed from Table 3(a)(b)(c) that the developed SVM model of Class 0, Class 1 and Class 2 with radial basis function has highest support vectors compared to other kernel function. Also SVM model Class 1 with radial basis function has highest number support vectors in the model.

c. **Langrangian multipliers(Alpha):** Alpha are the nonnegative Lagrange multipliers associated with the malaria optimization problem constraints $y_i[(w,x_i)]+b\geq 1$, $i=1,\ldots,n$.. It was observed from Table 3 (a) (b) (c) above that the developed SVM model of Class 0, Class 1 and Class 2 with radial basis function has high Langrangian Multipliers compared to other kernel function. Also SVM model Class 1 with radial basis function has highest number of Langrangian Multipliers in the model.

d. **Regularization Constant (C):** is the soft margin cost function of classification or penalty factor. A large C indicates low bias and high variance. Low bias because you penalize the cost of misclassification high **"hard margin"**. A small C indicates higher bias and lower variance and makes the cost of misclassification low "**soft margin**" the decision surface smooth. The cost function of the SVM model for Class 0, Class 1 and Class 2 irrespective of the kernel function used is 2×2 double vector.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

e. **MSE**: It was observed in Table 3 (a) (b) (c) that radial and polynomial kernel function gave the lowest mean square error of 0.5778 and 0.4444 respectively compared with linear kernel function. In the model polynomial function has highest bias of -3.7651 compared with radial basis function with lowest bias -0.0817.

f. **Accuracy**: It was observed in Table 3 (a) that polynomial kernel function has the highest accuracy of 88.89% but lowest support vectors, high bias and lowest alpha. Thus, the accuracy of the SVM model as shown in Figure 4.5 (a) (b) (c), Figure 4.6 (a) (b) (c) and Figure 4.7 (a) (b) (c) was derived from Class_2 with radial basis function of 85.60% accuracy and 308x8 double support vectors. This result indicates that performance analysis should not depend on accuracy alone but on other criteria.

g. False Positive Rate and False Negative Rate

Table 4.6 (a) (b) (c) shows the predicted result obtained and the target values for linear, radial basis and polynomial function. It was observed that:

i. With linear kernel function, out of 69 positive cases, 52 were identified positive while 17 were false negative. Thus class 2 linear function gave TP_R of 75.36%, TN_R of 81.89%, FP_R of 0.1802 and FN_R of 0.2464

ii. With radial basis function, out of 69 positive cases, 58 were identified positive while 11 were false negative. Thus class 2 radial basis function gave TP_R of 84.06% TN_R of 0.8649 FP_R of 0.1351 and FN_R of 0.1594.

iii. With polynomial kernel function out of 69 positive cases, 63 were identified positive while 6 were false negative. Thus class 2 polynomial kernel function gave TP_R of 91.30% TNR of 87.36%, FP_R of 0.1262 and FN_R of 0.0872.

Conclusions

In this paper, the multiclass feature selection was handled by SVM. The model was trained and tested with large and small large datasets and SVM handles them well regardless of their sizes but it uses predefined function to optimize well.

After several testing of the SVM model with different kernel functions, SVM_2 with radial basis function gave the best result with highest support vectors with 85.60% accuracy, 84.06% Sensitivity, 86.49% Specificity and 308 X 8 support vectors.

Thus, SVM explicitly control the trade-off between Complexity and error. SVM also minimizes upper bound generalization error compared to local training. Finally, it solves the problems of over-fitting by optimizing the model parameters to feature selection.

References

1. Depinay, J. M. O., Mbogo, C. M., Killeen, G., Knols, B., Beier, J., Carlson, J., ...& McKenzie, F. E. (2004). A simulation model of African Anopheles ecology and population dynamics for the analysis of malaria transmission. *Malaria Journal*, *3*(1), 29.

2. Randolph, S. E. (2008). Tick-borne disease systems. *Rev sci tech Off int Epiz*, *27*(2), 1-15.

3. Alonso, P. L., Brown, G., Arevalo-Herrera, M., Binka, F., Chitnis, C., Collins, F., ... & Mendis, K. (2011). A research agenda to underpin malaria eradication. *PLoS medicine*, *8*(1), e1000406.

4. Bousema, T., Okell, L., Felger, I., & Drakeley, C. (2014). Asymptomatic malaria infections: detectability, transmissibility and public health relevance. *Nature Reviews. Microbiology*, *12*(12), 833.

5. Onuwa, O. B. (2014). Fuzzy Expert System for Malaria Diagnosis.

6. Keeling, M. J., & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton University Press.

7. Namdev, N., Agrawal, S., & Silkari, S. (2015). Recent advancement in machine learning based internet traffic classification. *Procedia Computer Science*, *60*, 784-791.

8. Maina, E. M., Oboko, R. O., & Waiganjo, P. W. (2017). Using Machine Learning Techniques to Support Group Formation in an Online Collaborative Learning Environment. *International Journal of Intelligent Systems & Applications*, *9*(3).

9. Zia, T., Abbas, Q., & Akhtar, M. P. (2015). Evaluation of feature selection approaches for Urdu text categorization. *International Journal of Intelligent Systems and Applications*, *7*(6), 33.

10. Goswami, S., Chakrabarti, A., & Chakraborty, B. (2017). An efficient feature selection technique for clustering based on a new measure of feature importance. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-12.

11. Mohamad, M. S. (2004). Feature selection method using genetic algorithm for the classification of small and high dimension data. In *Proc. Int. Symp. Info. Com. Tech., 2004* (pp. 13-16).

12. V. Sindhwani, P. Bhattacharyya, Subrata Rakshit: Information theoretic feature crediting in multiclass support vector machines. First SIAM Int. Conf. on Data Mining, 2001

13. S. Dumais, J. Platt, D. Heckerman, M. Sahami: Inductive learning algorithms and representations for text categorization. Proc. 7th Int. Conf. on Information and Knowledge Management, pp. 148–155, 1998.

14. T. Joachims: Text categorization with support vector machines: learning with many relevant features. Proc. 10th ECML. LNCS vol. 1398, pp. 137– 142, 1998

15. Chapelle, O., & Keerthi, S. S. (2008, August). Multi-class feature selection with support vector machines. In *Proceedings of the American statistical association*.

16. Bi, J., Bennett, K., Embrechts, M., Breneman, C., & Song, M. (2003). Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, *3*(Mar), 1229-1243.

17. Zhu, J., Rosset, S., Tibshirani, R., & Hastie, T. J. (2004). 1-norm support vector machines. In *Advances in neural information processing systems* (pp. 49-56).

18. Ch, S., Sohani, S. K., Kumar, D., Malik, A., Chahar, B. R., Nema, A. K., ...& Dhiman, R. C. (2014). A support vector machine-firefly algorithm based predicting model to determine malaria transmission. *Neurocomputing*, *129*, 279-288.

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.