A Study of Classification and Quantification Methods for Identification and Counting Malaria Mosquitos

Xuetong Wang

A thesis in fulfilment of the requirements for the degree of

Master of Information Technology

School of Computer Science and Engineering

Faculty of Engineering

The University of New South Wales

August 2024

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: **Lastname**

First name: **Firstname** Other name/s: **Othernames**

Abbreviation for degree as given in the University calendar: **Master**

School: **School of Computer Science and Engineering** Faculty: **Faculty of Engineering**

Title: A Study of Classification and Quantification Methods for Identification and Counting Malaria Mosquitos

Abstract

Abstract

Declaration relating to disposition of project thesis/dissertation

Signature **Firstname Lastname** Witness Date **17 November, 2020**

FOR OFFICE USE ONLY Date of completion of requirements for Award

Originality Statement

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Xuetong Wang
17 November, 2020

Copyright Statement

I hereby grant the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

Xuetong Wang
17 November, 2020

Authenticity Statement

I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

Xuetong Wang
17 November, 2020

Abstract

Malaria, a mosquito-borne disease, affects over 250 million people worldwide each year, leading to more than 600,000 deaths. Effective controlling mosquito populations, particularly Anopheles species, is crucial for combating this disease. This study employs innovative optical sensors to collect wingbeat data from flying mosquitoes, facilitating the development of machine learning models for species classification and quantification. This study assesses the efficacy of using several learning techniques for mosquito identification using data collected from four mosquito species under controlled laboratory and semi-field conditions. Various quantification methods, including Classify and Count (CC) and Adjusted Classify and Count (ACC) [9], were evaluated to estimate the number of mosquitoes captured in traps. Experiments show that for this project, the DT classifier and ACC method achieved the best results in counting accuracy.

Acknowledgement

I would like to express my special thanks to my supervisor, Gustavo Batista, for providing me with this interesting topic and for his detailed and patient guidance. Without him, I would not have been able to successfully complete this project.

At the same time, I would like to thank my family for accompanying me and for their meticulous care and concern when I encountered difficulties.

Contents

List of Figures

List of Tables

Abbreviations

ACC	Adjusted Classify and Count
CC	Classify and Count
DT	Decision Tree
EMQ	Expectation Maximization for Quantification
GNB	Gaussian Naive Bayes
KNN	K-Nearest NeighborS
LDA	Linear Discriminant Analysis
LGBM	Light Gradient Boosting Machine
LR	Logistic Regression
MLP	Multi-Layer Perceptron
PCC	Probabilistic Classify and Count
RT	Random Forest
SVM	Support Vector Machine

Chapter 1

Introduction

Malaria is transmitted by mosquitoes of the Anopheles genus, including *An. gambiae*, *An. arabiensis*, and *An. funestus.* Through classification systems, control agents can detect areas with high densities of disease vectors or pests more quickly and accurately, enabling more effective control measures. Previous studies, such as that by Y. Hao et al., proposed using insect flight sound textures for automatic classification [1]. However, it was found that this approach is limited to insects with distinct frequency characteristics. Therefore, additional features and temporal information become crucial.

This project aims to classify and count malaria mosquitoes using the sensors proposed in the aforementioned study. Researchers at the University of New South Wales, funded by the Innovative Vector Control Consortium, recently conducted a research project collecting data on four mosquito species under laboratory and semi-field conditions. These four species include the primary malaria vectors (*An. gambiae*, *An. arabiensis*, and *An. funestus*) and *Culex quinquefasciatus*, a common mosquito species. This dataset was collected at the Ifakara Health Institute's facilities in Bagamoyo, Tanzania.

Therefore, we will use laboratory data and different machine learning models and compare the accuracy of these methods. We will use quantification methods to estimate the number of mosquitoes captured using sensor-equipped traps. Our methods will be applied in semi-field trials. Mosquitoes are released into a cage and captured using traps. We will compare the estimates provided by our methods with the actual number of mosquitoes captured.

Chapter 2

Literature Review

This project mainly uses nine different machine learning methods for comparison, and mainly compares three of them, namely Decision Tree (DT), Light Gradient Boosting Machine (LightGBM), and Support Vector Machine (SVM). It uses several different quantification methods for evaluation, namely Classify and Count (CC) [2], Adjusted Classify and Count (ACC) [3], Probabilistic Classify and Count (PCC) [4], and Expectation Maximization for Quantification (EMQ) [2]. The most appropriate model for this experiment can be obtained by comparing the accuracy of these methods.

The decision tree is a widely used algorithm in machine learning. It can identify and exploit the relationship between variables to handle complex information classification problems [5], and because it is easier to implement and understand than other classification algorithms. This is a widely used tool in various machine learning tasks. The decision tree can be intuitively represented as a tree structure similar to a flowchart, where each internal node is represented by a rectangle and the leaf nodes are represented by an ellipse [6]. In applications, decision trees need to be pruned or set to a maximum depth to prevent overfitting. At the same time, it should be noted that it is very sensitive to small changes in the data. Different data splits can lead to completely different tree structures, thus affecting the entire model.

As the fastest tree-based algorithm [7], LGBM supports distributed training and can handle extremely large data sets. When computed on large data sets, it is generally efficiency and scalability [8]. The main advantage of this algorithm is that it is a lightweight model that requires very little memory to compute and provides very accurate results. At the same time, because LightGBM has many hyperparameters, careful parameter tuning is often required to achieve the best results.

Support Vector Machine is a supervised learning algorithm that is widely used in various classification and regression tasks. Classification tasks can be performed in vector space or feature space in linear support vector machines [9]. The vector space is the space of the scatter plot containing the original features, which can represent the magnitude and direction information of the data. The feature space is the space of the scatter plot containing the features transformed by the kernel function, in which each data carries information about a single data point in the vector space. For linearly inseparable data, SVM can use the kernel trick to map the data into a high-dimensional feature space to find nonlinear boundaries. However, the kernel trick usually increases the number of dimensions and thus increases the computation time.

The remaining methods are Multi-Layer Perceptron(MLP), Logistic Regression(LR), Random Forest(RF), K-Nearest Neighbors(KNN), Gaussian Naive Bayes(GNB) and Linear Discriminant Analysis(LDA). The basic principle of MLP is gradient descent on an error function, which is suitable for regression and classification [10]. LR is a supervised machine learning algorithm. The goal is to map a function from the features of the data set to the target to predict the probability that a new example belongs to one of the target classes [11]. RF is a combination of tree predictors where each tree depends on the value of an independently sampled random vector. Classification accuracy can be significantly improved by increasing the number of trees in a forest and letting them vote for the most popular categories [12]. KNN assigns class labels to the most K nearest patterns in the data space. As a local method, the nearest neighbor technique is powerful when the data set is large and the dimensionality is low [13]. Naive Bayes is a probabilistic classifier based on Bayes' theorem, strong independence assumption and independent feature model. In

GNB, the continuous values of each class are distributed according to the normal distribution [14]. LDA consists of finding a projected hyperplane that minimizes the inter-class variance and maximizes the distance between class projected means. This hyperplane can be used for classification, dimensionality reduction, and explaining the importance of a given feature [15].

CC is often not the optimal choice. Even if a classifier performs well in classification tasks, it may exhibit bias in quantification tasks [2]. This is because it tends to minimize the number of false positives (FP) or false negatives (FN), but at the cost of increasing the quantity of the other type of error, leading to unbalanced results. This results in inaccurate estimates of class prevalence. However, CC plays an important role in quantification research, since it is the baseline for any reasonable quantification method to improve [2]. PCC is a variant of CC [2]. PCC utilizes a probabilistic classifier to generate posterior probabilities for each item and calculates the expected scores for each class, providing a more comprehensive measure of uncertainty. This contributes to improving performance when dealing with imbalances and uncertainties. However, the application of PCC is constrained by the limitations of probability calibration, especially when the training distribution differs from the unlabeled distribution, as the underlying assumption of probability calibration relies on the IID assumption [2].

ACC is also an improved method based on CC [2]. It utilizes a classifier to categorize unlabeled data, observes the scores for each category, and then estimates the prevalence of each category by solving a system of linear equations. By adjusting the results of CC and considering the classifier's tendency for misclassification, ACC is more adaptable to different distributions and uncertainties. However, a drawback of ACC is that its estimates may fall outside the [0,1] range, requiring clipping and rescaling. In certain scenarios, applying the 'softmax' function may offer a better solution [2]. ACC provides a more accurate estimation of category prevalence, particularly when dealing with distribution shifts and imbalanced data.

The aforementioned methods all possess an inductive nature, whereas EMQ is a transductive quantification method [2]. It is trained by observing certain characteristics of the unlabelled examples it needs to predict, rather than relying solely on the training set. This method, which updates class prevalence values and posterior probabilities through a mutually recursive approach, can enhance adaptability to specific test sets.

Chapter 3

Method

3.1 Data

In this project, there are four different classes: *An. gambiae s.s.*, *An. arabiensis*, *An. funestus*, and *Culex quinquefasciatus.* In the training set, the number of each class is 3000. In the test set, the number of *An. funestus* is 512, the number of *Culex quinquefasciatus* is 522, the number of *An. gambiae s.s.* is 600, and the number of *An. arabiensis* is 428. Clearly, the quantities of different classes in the training set are balanced, whereas in the test set they are imbalanced.

3.2 Data Preprocessing

The data for this project are based on results from semi-field trials conducted at the Ikakara Health Institute (IHI) facility in Bagamoyo, Tanzania. Six different sensors were used to collect training data. Since there may be differences in characteristics between different sensors, there may be adverse effects. Therefore, we group the training data by sensor, which can improve the accuracy and robustness of the model.

3.3 Cross-Validation

To improve the accuracy of model evaluation, cross-validation is used. Each time the data from one sensor is used as the test set and other data is used as the training set, the test is repeated and the average is taken as the final result.

3.4 Counting Accuracy formula

In this experiment, the count accuracy formula is used to compare the count results predicted by the model with the actual count results to quantify the accuracy of the model. The accuracy result will be between 0 and 1. The closer to 1, the more accurate the prediction.

Equation 3.1 is for calculate counting accuracy.

$$Counting\ Accuracy = (1 - \frac{|Estimated\ Number\ of\ Mosquitoes - Actual\ Number\ of\ Mosquites|}{Actual\ Number\ of\ Mosquitoes}) \times 100\% \tag{3.1}$$

3.5 Model Selection

3.5.1 Classification Accuracy

In this experiment, the same data was used to train 9 models, namely LGBM, MLP, LR, SVM, DT, RF, KNN, GNB and LDA. The average accuracy criterion and the difference are used as evaluation criteria, and cross-validation is used to evaluate the performance of each model. From simple models DT, LR to complex models LGBM and MLP, these different models have their own characteristics. By comparing the performance of these models, the most suitable model for this project can be found.

Table 3.1: Average accuracy and standard deviation of 9 models

model	**LGBM**	**MLP**	**LR**	**SVM**	**DT**	**RF**	**KNN**	**GNB**	**LDA**
mean acc	0.575	0.569	0.495	0.550	0.490	0.566	0.423	0.450	0.491
std	0.044	0.023	0.019	0.037	0.035	0.027	0.034	0.024	0.019

As shown in the figure, the LGBM model achieved the highest accuracy, followed by MLP, SVM and RF. However, the standard deviation of the LGBM model is relatively high, which shows that the LGBM model is not stable enough on different data. In comparison, LR and LDA have the best stability.

3.5.2 Classify and Count

For this project, we hope to find a model that can accurately estimate the proportion of talk for each category, rather than just accurately classify a single sample. Therefore, even if some classifiers achieve high overall accuracy, they may still exhibit significant errors in predicting class proportions. Hence, we tested the aforementioned classifiers and compared their accuracy in Classify and Count quantification.

Table 3.2: Average counting accuracy and standard deviation of 9 models

model	Counting Accuracy				
	An. arabiensis	*Culex*	*An. funestus*	*An. gambiae s.s.*	mean
LGBM	0.815	0.979	0.978	0.858	0.908
MLP	0.699	0.889	0.925	0.885	0.850
LR	0.65	0.779	0.801	0.930	0.790
SVM	0.827	0.830	0.972	0.97	0.901
DT	0.800	0.970	0.977	0.793	0.885
RF	0.791	0.887	0.950	0.854	0.871
KNN	0.889	0.449	0.929	0.510	0.694
GNB	0.455	0.232	0.698	0.608	0.498
LDA	0.720	0.819	0.830	0.929	0.825

It can be observed that the SVM and LGBM models performed well overall, particularly achieving high accuracy for the *An. arabiensis*, *An. funestus*, and *An. gambiae s.s.* species. For *An. arabiensis*, the KNN model had the highest accuracy at 88.9%, with SVM and LGBM also performing well at 82.7% and 81.5%, respectively. For *Culex*, the

LGBM model performed the best with an accuracy of 97.9%, followed by the DT and SVM models with accuracies of 97.0% and 83.0%, respectively. For *An. funestus*, the SVM model performed the best with an accuracy of 97.2%, and the LGBM and DT models also performed very well with accuracies of 97.8% and 97.7%, respectively. For *An. gambiae s.s.*, the SVM model again performed the best with an accuracy of 97.5%, followed by the LR and LDA models with accuracies of 93.0% and 92.9%, respectively. The consistent performance of SVM and LGBM across multiple species indicates that these models exhibit good stability and reliability when processing the data.

3.5.3 Model Selection

The accuracy of KNN and GNB in Classify and Count is very low, indicating that these models are not suitable for this project. SVM, LGBM, and DT show consistent performance across multiple categories, demonstrating good stability and reliability in handling the data. Therefore, these three models are selected for testing on the test set.

Chapter 4

Experimental Results

The above three well-performing models were tested on the test set. The accuracy of each category in CC, ACC, PCC, and EMQ, as well as the mean accuracy, were observed and compared.

4.1 LGBM classifier

Table 4.1: Average counting accuracy and standard deviation of LGBM classifier

Quantification Methods	Counting Accuracy				
	An. arabiensis	*Culex*	*An. funestus*	*An. gambiae s.s.*	mean
CC	0.956(0.18)	0.958(0.09)	0.498(0.08)	0.577(0.14)	0.747(0.07)
ACC	0.618(0.17)	0.806(0.11)	0.605(0.05)	0.855(0.13)	0.721(0.05)
PCC	0.764(0.23)	0.907(0.09)	0.439(0.08)	0.69(0.11)	0.723(0.08)
EMQ	0.925(0.28)	0.885(0.13)	0.232(0.1)	0.298(0.15)	0.585(0.12)

In identifying *Culex*, the CC method performs best with an accuracy of 91.7%. For identifying *An. gambiae s.s.*, the PCC method performs best with an accuracy of 82.2%. However, for identifying *An. funestus*, the EMQ method performs the worst with an accuracy of only 20.8%. In terms of average statistical accuracy, CC and PCC have the best results. In terms of stability, ACC has the most stable performance, followed by CC.

4.2 SVM classifier

Table 4.2: Average counting accuracy and standard deviation of SVM classifier

Quantification Methods	Counting Accuracy				
	An. arabiensis	*Culex*	*An. funestus*	*An. gambiae s.s.*	mean
CC	0.691(0.26)	0.87(0.15)	0.949(0.07)	0.937(0.06)	0.862(0.08)
ACC	0.945(0.17)	0.606(0.12)	0.994(0.06)	0.767(0.12)	0.828(0.11)
PCC	0.60(0.17)	0.809(0.08)	0.876(0.1)	0.987(0.07)	0.818(0.06)
EMQ	0.354(0.19)	0.572(0.08)	0.769(0.12)	0.891(0.1)	0.647(0.13)

Regarding Table 4.2, the CC method demonstrates the highest average accuracy at 86.2%. In contrast, the PCC method exhibits the lowest standard deviation at 64.7%. Among the species, *An. funestus* and *An. gambiae s.s.* show both high accuracy and stable performance across various methods. However, *An. arabiensis* achieves high accuracy only with the ACC method; its performance on other methods is subpar, characterized by significant deviation and insufficient stability.

4.3 Decision Tree classifier

Table 4.3: Average counting accuracy and standard deviation of DT classifier

Quantification Methods	Counting Accuracy				
	An. arabiensis	*Culex*	*An. funestus*	*An. gambiae s.s.*	mean
CC	0.930(0.09)	0.849(0.09)	0.707(0.08)	0.932(0.08)	0.854(0.03)
ACC	0.940(0.14)	0.943(0.06)	0.928(0.06)	0.907(0.08)	0.930(0.06)
PCC	0.930(0.09)	0.849(0.09)	0.707(0.08)	0.932(0.08)	0.854(0.03)
EMQ	0.930(0.09)	0.849(0.09)	0.707(0.08)	0.932(0.08)	0.854(0.03)

Table 4.3 demonstrates the very high stability of the model by showing that the average standard deviation of the Decision Tree classifier for each approach is quite low. Of them, the model's identification of An. arabiensis and An. gambiae s.s. yielded the greatest results, with counting accuracy over 90% in both cases. But it does a bad job of identifying An. funestus. With the exception of the ACC method, which has a counting accuracy of over 90%, the other methods only have an accuracy of roughly 70%.

Chapter 5

Conclusion

This project initially compared nine different models using the training set and selected the three with the best performance(LGBM, SVM, and DT) for further testing. The experimental results indicate that the decision tree (DT) classifier achieved the best overall performance. It excelled in the accuracy (ACC) method, attaining a final counting accuracy of up to 93%, while the results for the other three methods were also above 85%. Conversely, the LGBM model exhibited the lowest counting accuracy, particularly on the EMQ method, where it fell below 60%. Additionally, the high standard deviation across each method suggests that the model's stability is insufficient. The decision tree algorithm focuses on a single feature at a time, selecting the optimal feature and the best split point. It extracts the feature value at this optimal split and divides the domain into two subdomains, repeating this process until a specific subdomain shows a significantly higher concentration of certain categories compared to the others [16]. Consequently, the decision tree demonstrates strong adaptability to data distribution and noise, effectively identifying the variables that have the most significant impact on the results.

References

[1] Y. Hao, B. Campana, and E. Keogh, "Monitoring and mining animal sounds in visual space," *Journal of insect behavior*, vol. 26, pp. 466–493, 2013.

[2] A. Esuli, A. Fabris, A. Moreo, and F. Sebastiani, *Learning to quantify.* Springer Nature, 2023.

[3] A. Moreo and F. Sebastiani, "Re-assessing the "classify and count" quantification method," in *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43.* Springer, 2021, pp. 75–91.

[4] T. Schumacher, M. Strohmaier, and F. Lemmerich, "A comparative evaluation of quantification methods," *arXiv preprint arXiv:2103.03223*, 2021.

[5] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Computing and Applications*, vol. 23, pp. 2387–2403, 2013.

[6] A. Priyam, G. R. Abhijeeta, A. Rathee, and S. Srivastava, "Comparative analysis of decision tree classification algorithms," *International Journal of current engineering and technology*, vol. 3, no. 2, pp. 334–337, 2013.

[7] R. M. Aziz, M. F. Baluch, S. Patel, and A. H. Ganie, "Lgbm: a machine learning approach for ethereum fraud detection," *International Journal of Information Technology*, vol. 14, no. 7, pp. 3321–3331, 2022.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[9] S. Suthaharan and S. Suthaharan, "Support vector machine," *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207–235, 2016.

[10] M. Riedmiller and A. Lernen, "Multi layer perceptron," *Machine Learning Lab Special Lecture, University of Freiburg*, vol. 24, 2014.

[11] E. Bisong and E. Bisong, "Logistic regression," *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, pp. 243–250, 2019.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[13] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[14] E. K. Ampomah, G. Nyame, Z. Qin, P. C. Addo, E. O. Gyamfi, and M. Gyan, "Stock market prediction with gaussian naïve bayes machine learning algorithm," *Informatica*, vol. 45, no. 2, 2021.

[15] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," *Robust data mining*, pp. 27–33, 2013.

[16] L. Breiman, *Classification and regression trees.* Routledge, 2017.