Identifying *Aedes aegypti* Mosquitoes by Sensors and One-Class Classifiers

Vinicius M.A. Souza(✉)

Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, São Carlos, SP, Brazil
vsouza@icmc.usp.br

Abstract. Yellow fever, zika, and dengue are some examples of arboviruses transmitted to the humans by the *Aedes aegypti* mosquitoes. The efforts to curb the transmission of these viral diseases are focused on the vector control. However, without the knowledge of the exact location of the insects with a reduced time delay, the use of techniques as chemical control becomes costly and inefficient. Recently, an optical sensor was proposed to gather real-time information about the spatio-temporal distributions of insects, supporting different vector control techniques. In field conditions, the assumption of knowledge of all classes of the problem, it is hard to be fulfilled. For this reason, we address the problem of insect classification by one-class classifiers, where the learning is performed only with positive examples (target class). In our experiments, we identify *Aedes aegypti* mosquitos with an AUC = 0.87.

Keywords: Optical sensor · Insect classification · One-class classifiers

1 Introduction

The *Aedes aegypti* mosquito is one of the most important vectors of arboviruses that affect human health, including yellow fever, chikungunya, zika, Japanese encephalitis, and dengue. The viruses are passed on to humans through the bites of an infective female *Aedes* mosquito, which mainly acquires the virus while feeding on the blood of an infected person.

In May 2015, the Pan American Health Organization issued an alert regarding the first confirmed Zika virus infections in Brazil. Since this identification, the virus has spread rapidly throughout the America. The illness is usually mild with symptoms lasting for several days to a week after being bitten by an infected mosquito. However, Zika virus infection during pregnancy can cause a serious birth defect called microcephaly, as well as other severe fetal brain defects [1].

Dengue is the most important vector-borne viral disease of humans and likely more important than malaria globally in terms of morbidity and economic impact [2]. Studies estimate that 3.6 billion people living in areas of risk, with 390 million

V.M.A. Souza—The author thank the financial support of FAPESP (Grants #2011/17698-5, and #2015/16004-0).

C. Beltrán-Castañón et al. (Eds.): CIARP 2016, LNCS 10125, pp. 10–18, 2017.
DOI: 10.1007/978-3-319-52277-7_2

dengue infections per year globally, of which 96 million manifests clinically [3,4]. According to the World Health Organization, only 9 countries had experienced severe dengue epidemics before 1970. The disease is now endemic in more than 100 countries. In Latin America, the incidence and severity of this disease have increased rapidly in recent years. In 2015, 2.35 million cases of dengue were reported in the Americas alone, of which 10,200 cases were diagnosed as severe dengue causing 1,181 deaths [5].

Currently, no licensed vaccine against dengue infection is available, and the most advanced vaccine candidate did not meet expectations in a large trial [6]. Thus, the efforts to curb the transmission of these viral diseases are focused on the vector control in order to reduce the population of *Aedes aegypti*. There are many methods to insect control, as biological, genetic technology, environmental management and chemical control. However, without the knowledge of the exact location of the insects with a reduced time delay, the use of these techniques becomes costly and inefficient.

Recently, a new optical sensor was proposed as a tool to gather information about the spatio-temporal distributions of insects and to control disease vectors by the use of this sensor combined with an electronic trap [7]. The sensor captures insect flight information using a source light and automatically classifies the insects according to their species using machine learning algorithms. This sensor can provide real-time population estimates of insect species, supporting the effective use of traditional strategies to vector control.

The previous efforts related to insect classification by optical sensors have focused on multiclass classifiers, such as Support Vector Machines, k-Nearest Neighbors, Random Forest, Deep Neural Network, among others [7–10]. In multiclass classification, we have n predefined classes composed by the set of class labels $Y = \{y_1, y_2, \ldots, y_n\}$, where the main goal of a classifier is to assign the most probable class label $y_i \subset Y$ for an unknown example $\overrightarrow{x}$, where $\overrightarrow{x} \in \mathbb{R}^d$ is a feature vector with d dimensions. This procedure can be problematic when the example does not belong to any of predefined classes.

For the effective use of the sensor in field conditions, we note that the assumption of knowledge of all classes made by multiclass classifiers, it is hard to be fulfilled. For example, it is estimated that only the insects of the order *Diptera*, has more than 240,000 different species, where about 120,000 are cataloged [11]. Thus, it is impossible to conduct a comprehensive data collection that covers all possible species to build a classification model with all possible species. In practice, this means that there is a high probability of the sensor to deal with unknown species. In this case, a multiclass classifier will assign an incorrect class label to this insect, due the lack of data from other possible species.

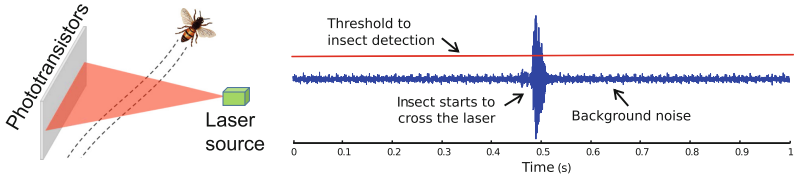
Given the need of identification of *Aedes aegypti* mosquitoes by sensors to support methods of vector control and the challenge to cope with unknown species, in this paper we address this classification problem using one-class classifiers [12]. In one-class classification, the learning is performed only with positive examples (target class) and none or few unlabeled examples from negative class.

We evaluated eight algorithms learned with only data from *Aedes aegypti* mosquitoes. The test was conducted with a dataset with five insect species collected by optical sensors. In our experimental evaluation, we conclude that the Parzen and SVDD are the most accurate algorithms for this application to the identification of *Aedes aegypti* mosquitoes.

The rest of the paper is organized as follows. Section 2 presents the optical sensor for insect classification. Section 3 describes the main concepts of one-class classification. Section 4 shows the results of our experimental evaluation. Finally, our conclusions are presented in Sect. 5.

2 Optical Sensor and Insect Data

The data evaluated in this paper was obtained from an optical sensor built with low-cost components to remotely capture information about flying insects. The sensor uses a light source, as a low-powered planar laser, that is pointed at an array of phototransistors as illustrated in Fig. 1-a). When a flying insect crosses the laser, its wings partially occlude the light, causing small variations in the light captured by the phototransistors. These variations are recorded as an audio signal, as the example presented in Fig. 1-b), given an *Aedes aegypti* crossing.



(a) Sensor (b) Signal

Fig. 1. Illustration of the optical sensor to capture information about insects and an example of audio signal collected given the crossing of an *Aedes aegypoti* mosquito.

In general, the data consist of background noise with occasional "events", resulting in the brief moment that an insect flies across the sensor. Note that the signal generated by the passage of the insect has an amplitude that is significantly higher than the amplitude of the background noise. In this way, using a simple threshold it is a trivial task to identify signal sections in which there is an insect passage. In contrast, the correct classification of each passage according to the insect species that generated the event is a more elaborate task. Basically, this task consists in extracting discriminant features from the signals for each species and using these features with machine learning algorithms.

2.1 Data Collection

In our study, we use the stream insect dataset previously evaluated in [10]. In this dataset, the collection was performed during six consecutive days in laboratory

Table 1. Insect dataset distribution.

Species of insect	Examples	Distribution (%)
Musca domestica	917	17.22
Culex quinquefasciatus	1,285	24.13
Culex tarsalis	1,265	23.76
Drosophila melanogaster	954	17.91
Aedes aegypti	904	16.98

conditions in which the temperature varied slightly between 20°C and 22°C and humidity varied between 20% and 35%. This dataset has insect passage signals from two species of flies and three species of mosquitoes. The flies species are the *Drosophila melanogasler* and the *Musca domestica*. The mosquito species are the *Culex quinquefascialus*, *Culex tarsalis* and the *Aedes aegypti*. It is interesting to note that *Culex* are species visually similar to *Aedes* and predominant in the Latin America houses. Table 1 presents a general description of the dataset.

2.2 Feature Extraction

In this work, we use the Mel-Frequency Cepstral Coefficients (MFCC) as recommended in a previous evaluation with a wide variety of signal processing techniques for feature extraction [7]. MFCCs are popular features in various application domains, particularly speech and speaker recognition [13].

MFCCs are calculated by taking the magnitudes of frequency components using an acoustically-defined scale called *mel* [14]. This scale relates physical frequencies to the frequencies perceived by the human auditory system. Equation 1 shows the conversion from frequency (f) to mel-frequency (m). Next, we apply a Discrete Cosine Transform. The MFCC are the cepstrum coefficients obtained from this operation. Specifically, we consider the 40 first coefficients as features.

$$m = 2595 \times log_{10}(1 + \frac{f}{700}) \tag{1}$$

3 One-Class Classification

Conventional multiclass classification algorithms aim to classify an unknown object into one of the several predefined categories. A problem arises when the unknown object does not belong to any of those categories. In one-class classification (OCC) [12,15], one of the classes (referred as target class) is well characterized by instances in the training data. For the other class (non-target), it has either no instances at all, very few of them, or they do not form a statistically representative sample of the negative concept.

In general, the problem of one-class classification is harder than the problem of conventional two-class or multiclass classification. For example, in binary

classification problems, the decision boundary is supported from both sides by examples of both classes. Because in the case of one-class classification only one set of data is available, only one side of the boundary is supported. It is therefore hard to decide, on the basis of just one class, how strictly the boundary should fit around the data in each of the feature directions [15]. This task is often called *data domain description*.

This OCC problem is often solved by estimating the target density or by fitting a model to the data support vector classifier. Instead of using a hyperplane, to distinguish between two classes, a hypersphere around the target set is used. The volume of the hypersphere is minimized directly. This method is called support vector data description (svdd) [16]. In svdd, a spherically shaped decision boundary around a set of objects is constructed by a set of support vectors describing the sphere boundary.

Different methods for data domain description have been developed. In this work, we evaluated eight different algorithms from the *Data Description toolbox* (DDtools) [12,17]. Specifically, the following algorithms: *gausdd* (Gaussian target distribution), *svdd* (support vector data description), *parzendd* (Parzen density estimator data description), *kmeansdd* (k-means data description), *knndd* (k-nearest neighbor data description), *lpdd* (linear programming data description), *mstdd* (minimum spanning tree data description), and *mogdd* (mixture of Gaussians data description). Unfortunately, due to space constraints, it is not possible to describe the algorithms. We direct the interested readers to [12] and [17] for a detailed explanation. However, an intuition of the decision boundary considered for each algorithm is shown in Fig. 2, given an artificial data example.

4 Experimental Evaluation

In our experimental evaluation, the classifiers were learned only with data of *Aedes aegypti* (target class). More specifically, we have considered the data from the first 48 h of the data collection, which represents 347 examples. To test the classifiers, we consider the remaining 557 examples from the class *Aedes aegypti* that was not used to train the classifiers and the 4,421 examples from the other four species of insects, totalizing 4,978 test examples.

Due to the imbalanced proportion of examples of target class compared to the non-target, a classifier that predicts the non-target class for all test examples achieves an accuracy around 90%. For this reason, we evaluate our results by the analysis of different performance measures, as Precision, Recall, F1-Score. Thus, given the rates of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) observed in a confusion matrix builded from the errors of a classifier, these measures are defined as follow:

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}, \quad F1 - Score = \frac{2\times(Precision\times Recall)}{Precision+Recall}$$

In addition, we also consider the measure Area Under Curve (AUC). This measure is related to the observed area on the Receiver Operating Characteristic

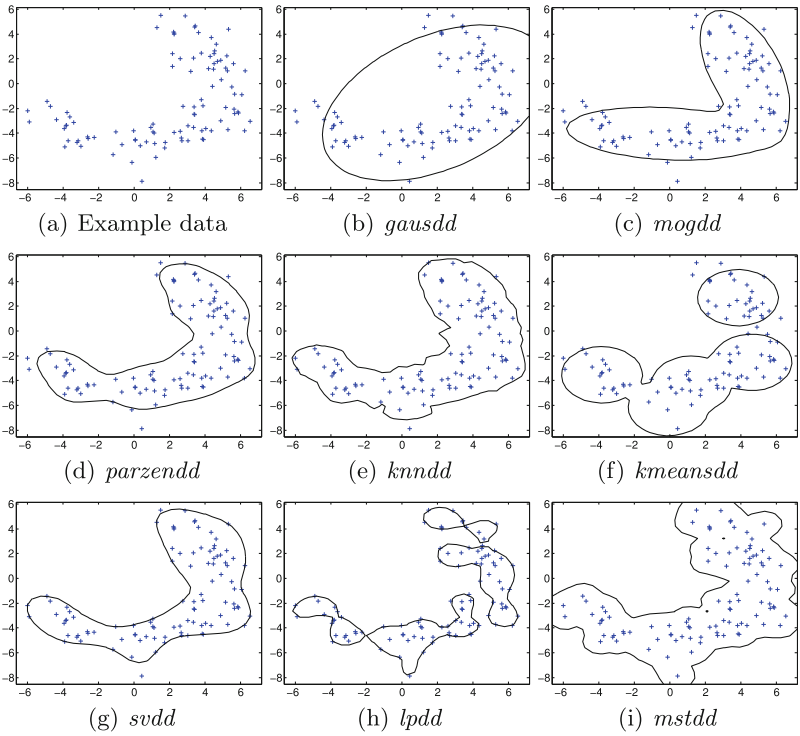


Fig. 2. Example of artificial bidimensional data and the decision boundary considered by each OCC algorithm evaluated.

curve (ROC curve). The ROC curve is a two-dimensional graphical representation which corresponds to false positive rate on the horizontal axis and the true positive rate on the vertical axis. Thus, in an ideal scenario, is expected a minimum value of false positives and a maximum value of true positives, which consequently leads to a value for AUC = 1.

The general results of the algorithms considering the five performance measures discussed are shown in Table 2. For each measure, the best result is highlighted. In this table, we also show the results achieved by a baseline which corresponds to a classifier that predicts the target class for all test examples.

We can see in Table 2 that the algorithm *parzendd* showed the best results for the measures Recall and AUC. On the other hand, the algorithm *svdd* showed the best results for the measures F1-Score and Accuracy. To better compare the results, the ROC curves achieved by the algorithms are shown in Fig. 3.

From the results showed in Table 2 and Fig. 3, we can note that both *parzendd* and *svdd* are very competitive, but the *svdd* showed results better balanced in terms of false positive and true positive rates. Although the *parzend* algorithm correctly identifies a higher number of *Aedes aegypti* mosquitoes, it also incorrectly identifies a higher number of insects from other species as *Aedes*. In Table 3 we shown more details about the errors of both algorithms.

Table 2. Results of one-class classifiers.

Algorithm	Precision	Recall	F1-Score	Accuracy	AUC
gausdd	0.45	0.76	0.57	86.96	0.82
mogdd	0.64	0.64	0.64	91.98	0.80
parzendd	0.41	**0.91**	0.56	84.35	**0.87**
knndd	0.43	0.87	0.57	85.42	0.86
kmeansdd	0.41	0.78	0.54	85.05	0.82
svdd	**0.78**	0.73	**0.75**	**94.62**	0.85
lpdd	0.32	0.85	0.46	77.90	0.81
mstdd	0.32	0.89	0.48	78.04	0.83
Baseline	0.10	1,00	0.18	10.98	0.50

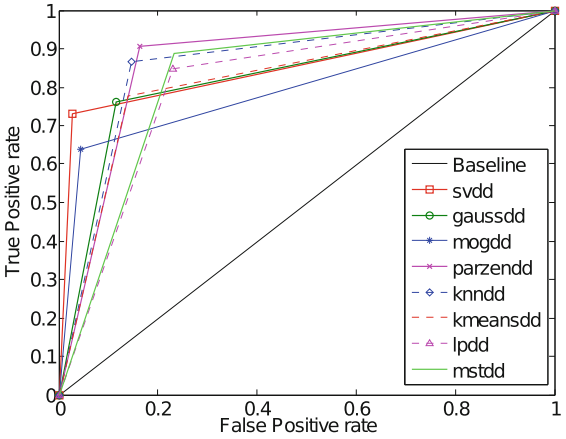


Fig. 3. The ROC curves of the OCC algorithms evaluated.

Table 3. Confusion matrices showed by the algorithms *parzendd* and *svdd*.

Actual	Predicted	
	Ae. aegypti	**¬ *Ae. aegypti***
Ae. aegypti	505	52
¬ *Ae. aegypti*	727	3694

parzendd

Actual	Predicted	
	Ae. aegypti	**¬ *Ae. aegypti***
Ae. aegypti	407	150
¬ *Ae. aegypti*	118	4303

svdd

5 Conclusion

In this paper, we showed an evaluation of one-class classifiers for the recognition of *Aedes aegypti* mosquitoes by optical sensors. *Aedes aegypti* is one of the most important vector of arboviruses as yellow fever, chikungunya, zika, and dengue. Thus, the recognition task is essential to support the efficient use of traditional methods to reduce the mosquitoes population, given the spatio-temporal

informations provided by the sensors. From the results, we conclude that even with a reduced number of target examples for training the classifiers (347 examples) and the absence of non-target examples, we can learn accurate classifiers. Among the evaluated algorithms, *svdd* and *parzendd* showed the best results, with AUC = 0.85 and AUC = 0.87, respectively. In future works, we want to explore the combination of different OCC algorithms and feature sets, and in conditions with concept drifts and extreme latency to update the classification model [18,19].

References

1. Plourde, A.R., Bloch, E.M.: A literature review of Zika virus. Emerg. Infect. Dis. **22**(7), 1185–1192 (2016)
2. Gubler, D.J.: The economic burden of dengue. Am. J. Trop. Med. Hyg. **86**(5), 743–744 (2012)
3. Beatty, M.E., Letson, G.W., Margolis, H.S.: Estimating the global burden of dengue. Am. J. Trop. Med. Hyg. **81**(5), 231 (2009)
4. Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O.: The global distribution and burden of dengue. Nature **496**(7446), 504–507 (2013)
5. W.H.O.: Dengue and severe dengue. Technical report Fact Sheet 117, World Health Organization (2015)
6. Halstead, S.B.: Dengue vaccine development: a 75% solution? The Lancet **380**(9853), 1535–1536 (2012)
7. Silva, D.F., Souza, V.M.A., Ellis, D.P.W., Keogh, E.J., Batista, G.E.A.P.A.: Exploring low cost laser sensors to identify flying insect species. J. Intell. Robot. Syst. **80**(1), 313–330 (2015)
8. Qi, Y., Cinar, G.T., Souza, V.M.A., Batista, G.E.A.P.A., Wang, Y., Principe, J.C.: Effective insect recognition using a stacked autoencoder with maximum correntropy criterion. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1–7 (2015)
9. Silva, D.F., Souza, V.M.A., Batista, G.E.A.P.A., Keogh, E., Ellis, D.P.W.: Applying machine learning and audio analysis techniques to insect recognition in intelligent traps. In: Proceedings of the International Conference on Machine Learning and Applications, pp. 99–10 (2013)
10. Souza, V.M.A., Silva, D.F., Batista, G.: Classification of data streams applied to insect recognition: initial results. In: Proceedings of the Brazilian Conference on Intelligent Systems, pp. 76–81 (2013)
11. Wiegmann, B., Yeates, D.K.: The Tree of Life Diptera (1996)
12. Tax, D.M.J.: One-class classification. Ph.D. thesis, TU Delft, Delft University of Technology (2001)
13. Zhen, B., Wu, X., Liu, Z., Chi, H.: On the importance of components of the MFCC in speech and speaker recognition. Acta Scietiarum Naturalium **37**(3), 371–378 (2001)
14. Stevens, S.S., Volkmann, J., Newman, E.B.: A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. **8**(3), 185–190 (1937)
15. Tax, D.M.J., Duin, R.P.: Uniform object generation for optimizing one-class classifiers. J. Mach. Learn. Res. **2**, 155–173 (2002)

16. Tax, D.M.J., Duin, R.P.W.: Support vector domain description. Pattern Recogn. Lett. **20**(11), 1191–1199 (1999)
17. Tax, D.: Ddtools, the data description toolbox for matlab version 2.1.2, June 2015
18. Souza, V.M.A., Silva, D.F., Gama, J., Batista, G.E.A.P.A.: Data stream classification guided by clustering on nonstationary environments and extreme verification latency. In: Proceedings of the SIAM International Conference on Data Mining, pp. 873–881 (2015)
19. Souza, V.M.A., Silva, D.F., Batista, G.E.A.P.A., Gama, J.: Classification of evolving data streams with totally delayed labels. In: Proceedings of the International Conference on Machine Learning & Applications, pp. 214–219 (2015)