Resultados Preliminares na Classificação de Insetos Utilizando Sensores Ópticos

Diego F. Silva[1], **Gustavo E. Batista**[1 2], **Eamonn Keogh**[2], **and Agenor Mafra-Neto**[3]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo

[2]Dept. of Computer Science & Engineering – University of California, Riverside

[3]ISCA Technologies, Riverside

diegofsilva@icmc.usp.br, gbatista@icmc.usp.br, eamonn@cs.ucr.edu,

president@iscatech.com

Abstract. *In this work we present a low-cost optical sensor to automatically count and classify disease vector insects in real time. We show that although the counting task is relatively straightforward, the insect classification in species is more elaborated and requires the identification of attributes in data. We evaluate two attributes: the wing-beat frequency and the circadian rhythm; as well as we present additional attributes that can be incorporated to the classifiers in future research. Our results are promising, with data collected with three insect species, we were able to classify them with accuracy higher than 90%.*

Resumo. *Neste trabalho nós apresentamos um sensor óptico de baixo custo para contagem e classificação automática de insetos vetores de doenças em tempo real. Nós mostramos que embora a tarefa de contagem seja relativamente simples, a classificação dos insetos em suas espécies é mais elaborada e requer a extração de atributos dos dados. São avaliados dois atributos: a frequência de batida de asas e o ritmo circadiano; bem como são apresentados outros atributos que podem ser incorporados aos classificadores em pesquisas futuras. Nossos resultados preliminares são promissores, com dados coletados com três espécies de insetos, foi possível classificá-los com acurácia acima de 90%.*

1. Introdução

É inegável a importância dos insetos para a vida humana, seja benéfica ou maleficamente. Por exemplo, insetos são vetores de doenças que matam dois milhões de pessoas por ano, deixando outras 700 milhões adoecidas [WHO 2010], além de causar perdas estimadas em dezenas de bilhões na agricultura e pecuária. Ao mesmo tempo, estima-se que insetos polinizam a maioria das espécies empregadas na agricultura, sendo que um terço de todo alimento consumido no mundo é polinizado somente por abelhas [Dixon 2009].

Dada a importância dos insetos para a vida humana, nas últimas décadas pesquisadores têm desenvolvido métodos de controle de insetos. Por exemplo, foram propostas inúmeras técnicas para controlar infestações de mosquitos responsáveis por transmitir

doenças como a malária e a dengue, tais como a pulverização de inseticidas e larvicidas; a introdução no ambiente de predadores de larvas e insetos adultos; a dispersão de mosquitos machos estéreis; a redução de *habitat* como a remoção de objetos que podem acumular água parada; o uso de armadilhas, entre outras técnicas [Walker 2002]. Para serem empregados, tais métodos de controle requerem o conhecimento da distribuição espaço-temporal dos insetos. Sem tal conhecimento, o emprego dessas técnicas se torna ineficiente.

Atualmente, estudar a distribuição espaço-temporal de insetos é uma tarefa custosa em termos de recursos e tempo de especialistas. Em linhas gerais, contagens de insetos são realizadas por meio de armadilhas, que são recolhidas periodicamente e analisadas por especialistas que identificam e contam manualmente as espécies. Além de ser uma abordagem cara, possui um atraso entre o momento que a armadilha é instalada e analisada, que pode ser de somente uma semana, mas esse tempo pode representar mais de meia vida de um inseto adulto. Dessa maneira, a doença já pode ter contaminado um grande número de pessoas no momento em que os dados estejam disponíveis aos especialistas [Patnaik et al. 2007]. Existe, portanto, uma necessidade por sensores automáticos e precisos, capazes de detectar, classificar e contar em tempo real insetos de diferentes espécies.

Estamos propondo um sensor que utiliza um feixe de luz laser para capturar dados de insetos à distância. Quando um inseto atravessa esse feixe, um fotosensor captura a variação da luz resultante da oclusão parcial do feixe causada pelo inseto. Essa variação da luz possui informações como a frequência de batida de asas do inseto, que podem ser utilizadas para classificar os insetos.

Apesar desse sensor poder ser aplicado a qualquer inseto com asas, estamos interessados na classificação de mosquitos vetores de doenças, em especial dos gêneros *Anopheles* e *Aedes*, vetores da malária e dengue, respectivamente. Estima-se que, por ano, a dengue afeta 50 milhões de pessoas e é considerada uma doença endêmica em mais de 100 países. A malária afeta entre 300 e 500 milhões anualmente, matando aproximadamente um milhão de pessoas.

Este trabalho mostra como os dados obtidos pelo sensor proposto podem ser utilizados para classificar insetos. Concretamente, nós mostramos resultados obtidos com três espécies de insetos, sendo uma espécie de abelhas e duas espécies de mosquitos. Para se obter uma alta acurácia de classificação é necessário identificar e extrair atributos dos dados e utilizar tais atributos para construir os classificadores. Tal argumento é sustentado por uma série de experimentos com algoritmos de classificação que utilizam os dados sem atributos previamente identificados e outros que utilizam atributos como a frequência de batida de asas do inseto e o ritmo circadiano. Ainda, apresentamos alguns atributos adicionais que podem ser extraídos e utilizados em pesquisas futuras.

Este trabalho está organizado da seguinte maneira: na Seção 2 é apresentado o sensor óptico para capturar dados de insetos; na Seção 3 é discutido como os dados de insetos foram coletados e posteriormente utilizados para construir os classificadores; na Seção 4 é realizada uma discussão de quais atributos adicionais podem ser utilizados para classificar os insetos; por fim, na Seção 5 são apresentadas as conclusões deste trabalho.

2. Sensor Classificador de Insetos

Estamos desenvolvendo o hardware e o software de um sensor para capturar informações de insetos à distância. A ideia central desse sensor é utilizar um feixe de luz laser, um refletor total e um fotosensor, como mostrado na Figura 1. Quando um inseto atravessa o feixe de luz, o fotosensor captura a variação resultante da oclusão parcial da luz pelo inseto. Essa variação é registrada na forma de uma série temporal, a qual armazena informações que podem ser utilizadas na classificação. O objetivo é construir um sistema que utilize tais séries temporais como entrada e forneça contagens dos insetos separadas por espécie e sexo, uma vez que somente as fêmeas de mosquitos são vetores de doenças.

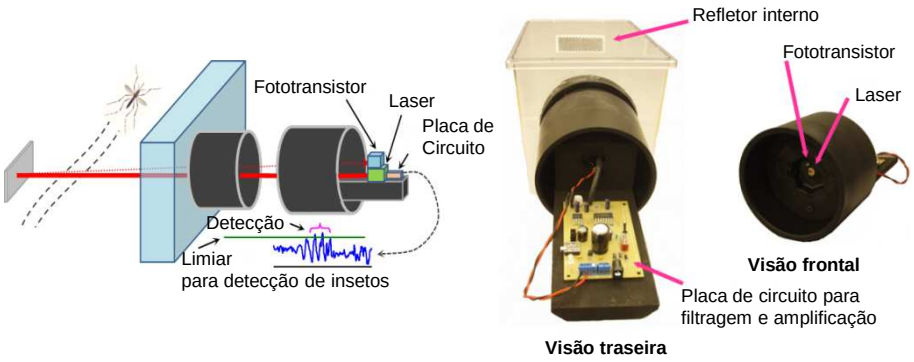

Figura 1. Projeto lógico do sensor em desenvolvimento (esquerda) e uma fotografia do sensor em sua versão atual (direita)

O sensor utiliza um circuito especialmente projetado para filtrar e amplificar os sinais capturados pelo fototransistor, atualmente gravados como arquivos de áudio. Apesar dos dados serem obtidos opticamente, os sinais soam exatamente como o som dos insetos capturados por microfones, com duas vantagens importantes: a primeira é que a luz laser pode trafegar grandes distâncias sem perda significativa de intensidade e, portanto, o sensor pode cobrir uma área de pelo menos algumas dezenas de metros quadrados; a segunda é que o sensor é totalmente *surdo* para qualquer outra interferência que não atravesse a luz laser, como a voz de pessoas, pássaros e aviões, portanto é menos susceptível a falsos positivos.

Nosso objetivo é que haja pelo menos alguns milhares desses sensores espalhados em campo e em países afetados por doenças como a malária e a dengue. Sendo assim, o sensor deve ser de baixíssimo custo, utilizando componentes facilmente encontrados. Idealmente, uma unidade do sensor deve custar menos de R$10,00. Dessa forma, a utilização ampla se torna mais fácil e, por ter baixo valor comercial e de troca, desestimula a prática de roubo.

3. Avaliação Experimental

Iniciamos esta seção explicando a nossa filosofia experimental. Nós projetamos todos os experimentos de forma que eles sejam *facilmente* reproduzíveis. Por exemplo, se o leitor quiser reproduzir qualquer figura deste artigo, ele pode simplesmente executar um programa para essa tarefa. Para isso, foi criado um sítio web[1] com todos os códigos-fonte

[1] `http://www.icmc.usp.br/~diegofsilva/ENIA2011`

de programas e conjuntos de dados, bem como planilhas com resultados experimentais detalhados. Todos os programas estão disponíveis em Matlab, o qual possui um clone totalmente gratuito (Octave) disponível. Além disso, disponibilizamos todos os arquivos de projeto de hardware e software do sensor em um sítio web[2] e estamos distribuindo gratuitamente hardware e software do sensor para pesquisadores em Entomologia interessados em construir uma base de dados com sinais de insetos de diferentes espécies.

A seguir, descrevemos como os dados foram coletados e pré-processados utilizando três espécies de insetos e como os classificadores foram costruídos a partir desses dados. Acreditamos que o grande desafio de construir classificadores precisos neste domínio está na busca por atributos. Para comprovar essa hipótese, iniciamos nossos experimentos utilizando diversos paradigmas de aprendizado para aprender diretamente com os dados coletados como exemplos. Diante dos resultados modestos obtidos com os dados originais, investigamos a extração de atributos, particularmente a frequência de batida de asas e o ritmo circadiano dos mosquitos.

3.1. Coleta de Dados

Os dados foram coletados durante 15 dias de três espécies de insetos: *Bombus impatiens*, espécie de abelhas, benéfica ao ser humano; *Aedes aegypti*, vetor da dengue e da febre amarela; e *Culex quinquefasciatus*, vetor de diversas doenças graves como encefalites (em especial St. Louis e Venezuelana), filariose bancroftiana e febre do Nilo Ocidental.

Os dados foram coletados em laboratório, com condições de temperatura e umidade controladas. As temperaturas durante as coletas variaram de 21^{o}C a 23^{o}C e a umidade entre 50% e 70%. Os arquivos de áudio possuem uma taxa de amostragem de 44100Hz, sendo posteriormente amostrados para 16000Hz, a fim de reduzir os requisitos de armazenamento e processamento dos dados. Essa taxa de amostragem é suficiente para caracterizar os sinais de insetos, uma vez que é capaz de representar até frequências de 8000Hz, sendo que os insetos normalmente apresentam sinais na faixa de 100Hz a 1000Hz.

Os arquivos de áudio consistem em geral de ruído de fundo com "bips" ocasionais, resultado dos breves cruzamentos do inseto com o laser. Na Figura 2, é exibido um exemplo do dado coletado pelo sensor. O inseto em questão é da espécie *Aedes aegypti*. Note que o sinal gerado pelo inseto possui uma amplitude significativamente maior que a amplitude do ruído de fundo. Dessa maneira, é uma tarefa relativamente simples *contar* o número de insetos que cruzam o laser, sendo que *classificar* os sinais nas espécies é uma tarefa bem mais sofisticada.

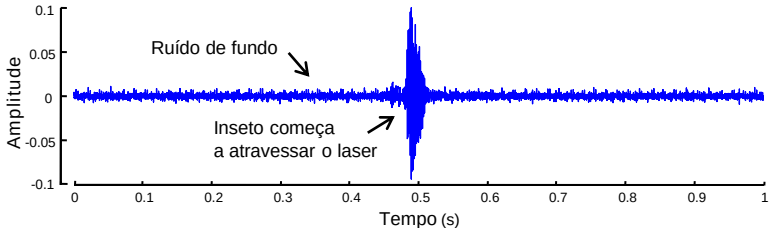

Figura 2. Exemplo do sinal gravado pelo sensor, espécie *Aedes aegypti*

Para separar o ruído de fundo dos sinais gerados pelos insetos, foi implementado

[2] http://www.cs.usr.edu/~eamonn/CE

um detector, descrito no Algoritmo 1. Esse algoritmo utiliza duas sub-rotinas auxiliares: normalizaZ que realiza normalização em índices z, e TRF que calcula a transformada de Fourier. O detector simplesmente utiliza uma janela deslizante sobre os dados, e calcula o espectro do sinal dentro da janela. Como a maioria dos insetos possui frequência de batida de asas na faixa entre 100Hz e 1000Hz, utilizamos a magnitude máxima no espectro do sinal dentro dessa faixa de frequências como um valor de confiança para o detector. Dessa maneira, quanto maior for a magnitude do sinal nessa faixa, maior a confiança de que o sinal não é um ruído de fundo. Todos os sinais com magnitude acima de um limiar especificado pelo usuário são considerados um evento gerado por um inseto. A alta razão sinal-ruído dos dados coletados pelo sensor permite ao usuário especificar valores baixos para o limiar de forma a garantir a identificação de sinais curtos ou de baixa amplitude, sem o risco de falsos positivos. Na Figura 3, ilustramos como o detector funciona.

Algoritmo 1: Detector da passagem do mosquito pelo sensor

Entrada: Um vetor com dados de áudio a
Um limiar de aceitação l
Um tamanho do passo t_p
Uma frequência mínima de interesse min_f = 100 Hz
Uma frequência máxima de interesse max_f = 1000 Hz
início
  a_z = normalizaZ(a)
  para $i = 0$ *até tamanho*(a_z)$-t_p$, *passo* t_p **faça**
    $janela = a_z(i : i + t_p)$
    f = TRF($janela$)
    $d(i : i + ws) = max(abs(f(min_f : max_f)))$
  fim para
  retorna d
fim

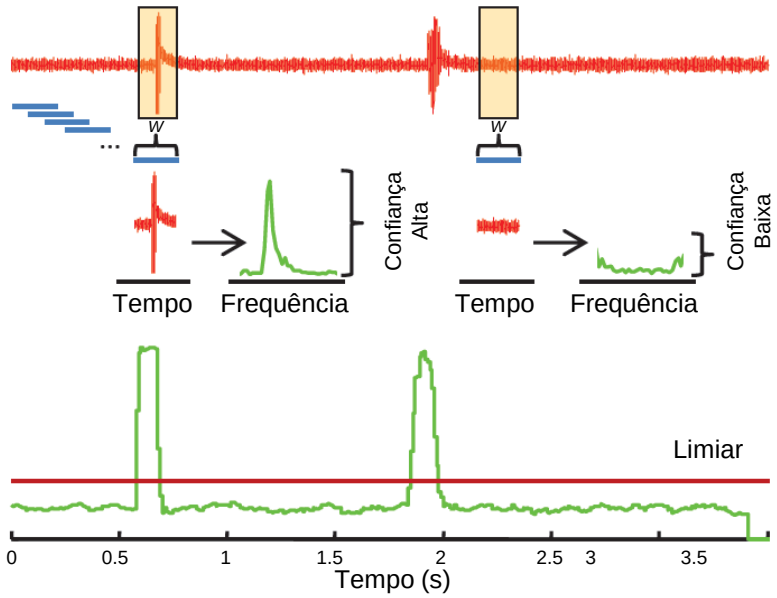

Figura 3. Esquema do detector de batida de asas

O passo final de pré-processamento consiste em extrair fragmentos com eventos

individuais utilizando o detector. Como os eventos gerados pelos insetos possuem durações diferentes, decidimos extrair fragmentos de um segundo. Tais fragmentos são longos o suficiente para armazenar a maioria dos eventos, os quais tipicamente duram um ou dois décimos de segundo. Note que os eventos podem ocorrer com qualquer intervalo de tempo entre eles, de forma que não é possível garantir que cada fragmento de um segundo possui somente um único evento. Entretanto, podemos garantir que cada fragmento possui ao menos um evento. Uma descrição resumida dos dados coletados pelo sensor é mostrada na Tabela 1.

Tabela 1. Descrição resumida dos dados coletados

Total de Amostras	Classe	Exemplos	Distribuição (%)
5982	*Aedes Aegypti*	1231	20,58%
	Bombus impatiens	499	8,34%
	Culex quinquefasciatus	4252	71,08%

3.2. Classificação sem a Identificação de Atributos

Como descrito na Tabela 1, a detecção e pré-processamento dos dados resultou em 5982 séries temporais, sendo que cada uma dessas séries tem duração de um segundo, ou 16000 observações. Uma forma bastante simples de interpretar esses dados é pensá-los como uma tabela atributo-valor com 5982 exemplos e 16000 "atributos". Cada exemplo foi rotulado de acordo com o inseto que gerou os dados, uma vez que os dados foram coletados com cada espécie de inseto separadamente. As limitações dessa representação para este problema são bastante evidentes. Por exemplo, apesar das séries temporais representarem áudios de um segundo, a duração dos sinais é variável, tendo duração típica de um a dois décimos de segundo (vide Figura 2). Por outro lado, tal representação permite a utilização de um grande número de algoritmos de classificação disponíveis e tais algoritmos podem prover resultados que podem ser utilizados como base para comparação para outras técnicas a serem avaliadas.

Neste experimento, utilizamos a ferramenta de domínio público Weka [Witten and Frank 2005]. Para a execução dos experimentos, foram utilizados quatro diferentes algoritmos de classificação: IBk (implementação do método k-vizinhos mais próximos, com os parâmetros k=5 e votação ponderada pelo inverso da distância); classificador Naïve Bayes; J48 (implementação do algoritmo de árvore de decisão C4.5); e SMO (implementação de uma Máquina de Suporte Vetorial). Todos os experimentos foram feitos utilizando o método de *resampling 10-fold cross validation*. Nós reportamos a acurácia média nos 10 conjuntos de teste e os respectivos desvios padrão como as nossas principais medidas de desempenho. Os resultados são exibidos na Tabela 2.

Tabela 2. Descrição resumida dos resultados da classificação utilizando o sinal não processado

Algoritmo	Acurácia	Desvio Padrão
IBk	68,34%	1,58
J48	66,50%	1,86
Naïve Bayes	40,17%	1,72
SMO	67,34%	1,16

Nesses experimentos, o classificador que obteve a melhor acurácia foi o IBk com 68,34%, inferior à porcentagem de exemplos da classe majoritária, a qual é de 71,08%

(vide Tabela 1). Isso significa que esse método possui desempenho inferior ao classificador trivial que classifica todos os exemplos como pertencentes à classe majoritária.

Uma possível crítica a esses experimentos é que os sinais a serem classificados são muito mais curtos do que as séries temporais utilizadas como exemplos. Uma parte significativa das observações é meramente ruído de fundo, uma característica que pode dificultar, por exemplo, a interpretação de distâncias calculadas pelo classificador IBk. Entretanto, é importante ressaltar que o problema de separar sinal e ruído de fundo não é trivial. Isso ocorre, pois a transição entre ruído de fundo e sinal não é abrupta e a duração dos sinais, como observado anteriormente, é variável. Em casos especiais, alguns sinais tem duração próxima a um segundo, como no caso das abelhas que ocasionalmente pairavam em frente ao laser.

Para eliminar a possibilidade de que os resultados modestos apresentados pelos classificadores tenham como origem a curta duração média dos sinais em comparação com a duração total da série temporal, decidimos extrair séries temporais mais curtas, com duração de 3 décimos de segundo. Novamente, os dados foram transformados em uma tabela atributo-valor. Utilizamos o classificador IBk, dado que esse obteve a melhor acurácia nos experimentos anteriores. A acurácia obtida neste novo experimento foi de 70,49%, ou seja, ainda inferior à acurácia do classificador trivial.

Os resultados iniciais evidenciam que o principal desafio está em identificar atributos que permitam construir classificadores precisos. Nas próximas duas seções são mostrados dois exemplos concretos desses atributos.

3.3. Classificação Utilizando a Frequência de Batida de Asas

A frequência de batida de asas é um importante atributo para classificar insetos, o qual está relacionado com o tamanho dos insetos, sendo que insetos maiores tendem a apresentar frequências menores de batida de asas. No caso de mosquitos, esse atributo pode também ajudar a separar mosquitos machos e fêmeas em espécies dimórficas, sendo que as fêmeas tendem a ser maiores do que os machos e portanto apresentam frequências de batida de asas mais baixas.

A Figura 4-*a* mostra o espectro do sinal da Figura 2. Pode-se observar, além do pico na frequência de batida de asas do mosquito (frequência fundamental), as 3 harmônicas em frequências múltiplas inteiras da frequência fundamental.

Neste trabalho, utilizamos o conceito de *cepstro*, proveniente da áera de processamento de sinais, para identificar a frequência de batida de asas. Informalmente, o cepstro é capaz de identificar frequências harmônicas representá-las em um único valor. A Figura 4-*b* apresenta o cepstro do mesmo sinal. O maior pico está localizado na frequência de 0,001563s, aproximadamente 639,79Hz.

Os dados utilizados são os mesmos cuja coleta foi descrita na Seção 3.1. Foi calculado o cepstro para cada fragmento de um segundo e o valor máximo do cepstro foi anotado como a frequência da batida de asas do inseto. Na Figura 5 é mostrado um histograma criado a partir dos resultados obtidos com as três espécies. É possível perceber que as classes se assemelham a uma distribuição Gaussiana com uma longa cauda esquerda. Também pode-se notar que que as frequências da espécie *Bombus Impatiens* (abelhas) são linearmente separáveis das espécies de mosquitos. Analisando somente os mosquitos, *Aedes*

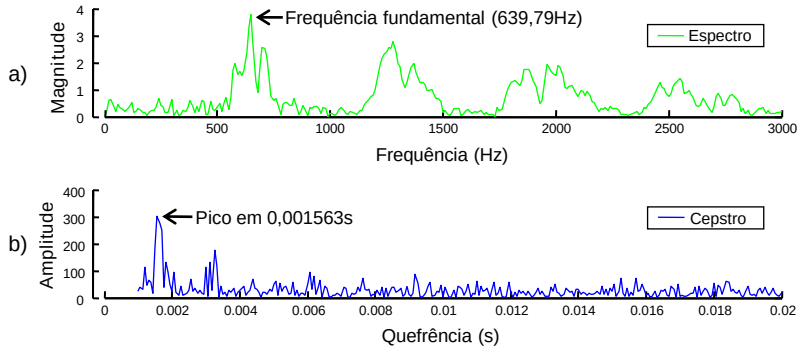

Figura 4. Espectro (a) e cepstro (b) do sinal obtido pelo sensor, que são utilizados para calcular a frequência da batida de asas de um inseto

aegypti possui uma frequência de batida de asas mais alta que o *Culex quinquefasciatus*. Entretanto, existe uma sobreposição nas frequências de batida de asas dessas duas espécies. A Tabela 3 provê algumas estatísticas para as frequências de batida de asas.

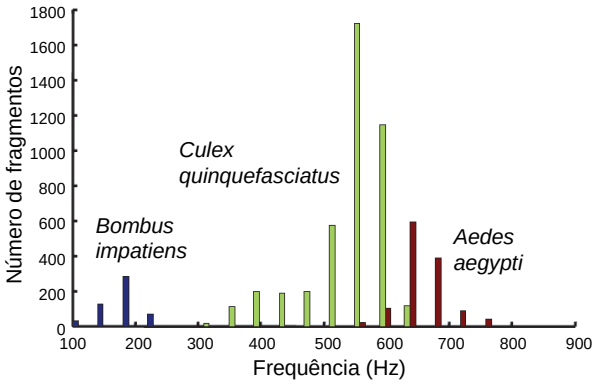

Figura 5. Histograma da frequência da batida de asas das espécies *Bombus impatiens*, *Aedes aegypti* e *Culex quinquefasciatus*

Tabela 3. Características dos dados de batida de asas

Classe	Exemplos	Média (Hz)	Desvio Padrão (Hz)
Aedes aegypti	1231	644,76	34,86
Bombus impatiens	499	173,76	26,42
Culex quinquefasciatus	4252	528,12	62,67

Nós podemos utilizar um classificador Bayesiano combinado com o critério de máximo a posteriori (MAP). A idéia geral do critério MAP é classificar um inseto, com uma dada frequência de batida de asas v, na classe de maior probabilidade a posteriori:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c|f = v)$$

no qual, C é o conjunto de todas as classes, f é uma variável aleatória para a frequência de batida de asas e v é o valor medido da frequência de batida de asas para o caso de teste. É possível utilizar a regra de Bayes para calcular a probabilidade a posteriori de cada classe, $P(c|f = v)$:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} \frac{P(f = v|c)P(c)}{P(f = v)}$$

O termo $P(f = v)$ pode ser descartado, pois ele é uma constante independente de c, resultando em:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(f = v|c)P(c)$$

Nós vamos assumir que as frequências de batida de asas são normalmente distribuídas para cada espécie. Assim, é possível estimar a quantidade $P(f = v|c)$ dada a equação da curva Gaussiana:

$$P(f = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

no qual, μ_c e σ_c^2 são as médias e variâncias das frequências de batida de asas para cada classe c. Esses parâmetros, bem como as probabilidades a priori das classes, $P(c)$, podem ser estimadas a partir dos dados, como mostrado na Tabela 3.

Ao aplicar o classificador Bayesiano MAP aos dados, foi obtida uma acurácia respeitável de 96.04%. Os resultados para cada classe podem são apresentados na Tabela 4.

Tabela 4. Sumário do desempenho do classificador Bayesiano para cada classe

	Predito			Acurária na classe
	Bombus impatiens	*Culex quinquefasciatus*	*Aedes aegypti*	
	Bombus impatiens	499	0	0	100,00%
Real	*Culex quinquefasciatus*	0	4139	113	97,34%
	Aedes Aegypti	0	124	1107	89,92%

Os resultados obtidos com a frequência de batida de asas são bastante promissores. Estamos coletando dados para mais espécies de insetos e ao considerar tais espécies adicionais, novos atributos serão necessários para manter uma alta acurácia. Na próxima seção, exploramos o fato de que insetos possuem períodos distintos de atividade como atributo adicional para o classificador.

3.4. Ritmo Circadiano

Muitas espécies de mosquitos são mais ativas durante determinados períodos do dia. Essa atividade é conhecida como ritmo circadiano, fato conhecido desde a antiguidade e que tem sido estudado por quase cem anos [Roubaud 1918].

Nós utilizamos o sensor para medir o ritmo circadiano de mosquitos, conforme é apresentado na Figura 6 para as espécies *Culex quinquefasciatus* e *Aedes aegypti*. Durante cinco dias, foi medida a atividade dos mosquitos em uma sala com iluminação natural. Para a espécie *Culex quinquefasciatus* foram coletados mais de 200.000 eventos e para a espécie *Aedes aegypti* cerca de 30.000 eventos. Os ritmos circadianos mostram que as duas espécies possuem padrões de atividade bastante distintos. Os mosquitos da espécie *Culex*

quinquefasciatus são predominantemente noturnos com um pico de atividade vespertina; os mosquitos da espécie *Aedes aegypti* são diurnos, também com um pico de atividade verpertina.

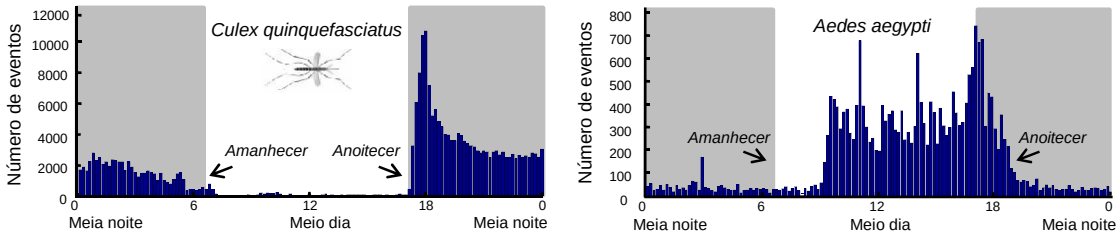

Figura 6. Ritmos cicadianos dos mosquitos das espécies *Culex quiquefasciatus* (*esquerda*) e *Aedes aegypti* (*direita*)

A Figura 7 fornece uma intuição de como é possível adicionar os ritmos circadianos à informação de batida de asas. Suponha um problema de classificar um inseto como *Aedes aegypti*, *Culex quinquefasciatus* ou *Anopheles stephensi* e que foi medida uma frequência de batida de asas a 428Hz. Utilizando somente essa informação, o inseto é igualmente provável de pertencer à classe *Anopheles stephensi* ou *Culex quinquefasciatus*. Entretanto, se o inseto foi observado às 11 horas, podemos verificar que é bem mais provável que seja um *Anopheles stephensi*.

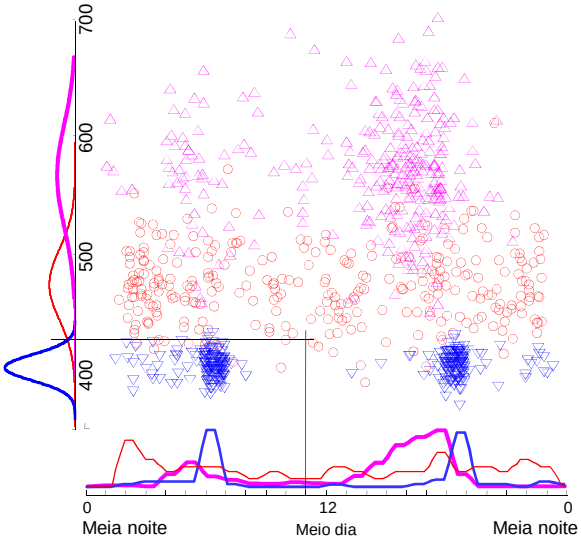

Figura 7. Gráfico de dispersão da frequência de batida de asas vs. horário de observação de 300 exemplos de *Aedes aegypti* △, *Culex quinquefasciatus* ▽ e *Anopheles stephensi* ○. Um inseto observado às 11 horas com frequência de batida de asas de 428Hz é quase certo ser um *Anopheles stephensi*

Atualmente, estamos coletando dados com informação temporal e de frequência de batida de asas para diversas espécies de interesse. Entretanto, podemos fazer um exercício para avaliar como os algoritmos de aprendizado de máquina estendidos com ambos os atributos poderiam beneficiar o aprendizado. Nós utilizamos dados gerados a partir de distribuições reais disponíveis na literatura para frequência de batida de asas [Reed et al. 1942] e ritmos circadianos [Taylor 1998] para as espécies *Aedes aegypti*, *Culex quinquefasciatus* e *Anopheles stephensi* (vide Figura 7).

Os experimentos utilizam 1000 exemplos, dois algoritmos de aprendizado de máquina (Naïve Bayes e k-vizinhos mais próximos) e três combinações de atributos. Na primeira, somente com a frequência de batida de asas foi utilizada no aprendizado, na segunda, apenas o ritmo circadiano e na terceira, ambos os atributos. Em todos os experimentos, foi utilizada a acurácia obtida com *leaving-one-out cross-validation* como principal medida para avaliar os resultados, apresentados na Tabela 5.

Os classificadores aprendidos a partir dos ritmos circadianos apresentam as acurácias mais baixas. Esse é um resultado esperado, uma vez que existe um alto grau de sobreposição nos períodos de atividade das três espécies (vide Figura 7). Os classificadores aprendidos com a frequência de batida de asas obtiveram acurácias consideravelmente mais altas. No experimento final, ambos os classificadores foram beneficiados pelo uso dos dois atributos combinados, com o Naïve Bayes sendo o classificador que obteve a maior melhora de desempenho, 93,18%. Essa é uma acurácia respeitável, considerando que estamos classificando espécies de mosquitos com características físicas e de comportamento similares, as quais até mesmo entomologistas têm dificuldades em distinguir.

Tabela 5. Acurácia para os classificadores k-vizinhos mais próximos e naïve bayes

Atributos	Naïve Bayes	k-NN
Ritmo circadiano	70,69%	68,50%
Frequência de batida de asas	90,73%	91,30%
Atributos combinados	93,18%	91,80%

4. Atributos Complementares

Nesta seção, são discutidos alguns atributos adicionais que podem ser utilizados em classificadores com o intuito de distinguir um número maior de espécies.

4.1. Características Meteorológicas

Temperatura, pressão atmosférica e umidade são as principais variáveis meteorológicas que afetam os insetos, por dois motivos principais. Em primeiro lugar, algumas espécies são mais adaptadas para viver em determinadas condições ambientais. Por exemplo, muitas espécies de mosquitos são originárias de regiões tropicais e subtropicais, onde o clima é quente e úmido. Portanto, em períodos de clima nessas condições, a prevalência de insetos dessas espécies pode aumentar. A segunda razão é que pelo menos a temperatura influencia o metabolismo dos insetos e as propriedades aerodinâmicas do ar. Em particular, esperamos que a frequência de batida de asas aumente proporcionalmente à temperatura. Há pesquisas significativas que sugerem que a temperatura tem efeito linear em uma ampla faixa de valores para a maioria dos insetos [Reed et al. 1942].

4.2. Características Intrínsecas

A frequência não é a única característica que pode ser retirada do sinal coletado pelo sensor. Insetos têm diferenças anatômicas que podem causar pequenas variações na forma das séries temporais coletadas pelo nosso equipamento. Por exemplo, a maioria dos insetos têm dois pares de asas, mas alguns deles têm apenas um par. Ainda, outras espécies têm um segundo par asas muito pequeno, usado principalmente para a estabilização de voo, como é o caso das moscas domésticas. Além disso, existem centenas de recursos de áudio que podem ser extraídos de arquivos de som. Um exemplo é o uso da magnitude das harmônicas de segunda ordem e superior como atributos [Moore and Miller 2002].

4.3. Características Espaço-temporais

Nossos sensores devem ser colocados em algum lugar no espaço e no tempo e esses dados espaço-temporais oferecem atributos potencialmente úteis [Grüebler et al. 2008]. As características espaço-temporais alteram a probabilidade a priori de uma determinada espécie, e essas informações podem ser inseridas no modelo de classificação. As características espaciais incluem altitude, distância a uma fonte de água doce, tipo de solo, velocidade média do vento, densidade populacional humana / insetos, tipo de atividade agrícola local, etc. As características temporais incluem padrões de época do ano e hora do dia (circadiano), que, como discutido anteriormente, serão particularmente interessantes.

5. Conclusões

Neste artigo, apresentamos um novo sensor para contar e classificar insetos à distância utilizando laseres. Mostramos que tal sensor pode facilmente contar insetos e que a tarefa de classificação é factível por meio da extração de atributos a partir dos dados extraídos pelo sensor. Concretamente, demonstramos o uso de dois atributos: a frequência de batida de asas e o ritmo circadiano; e ainda apresentamos outros atributos que podem ser incorporados aos classificadores em pesquisas futuras.

Referências

Dixon, K. W. (2009). Pollination and Restoration. *Science*, 325(5940):571–573.

Grüebler, M. U., Morand, M., and Naef-Daenzer, B. (2008). *Agriculture, Ecosystems and Environment*, 123(1–3):75–80.

Moore, A. and Miller, R. H. (2002). Automated identification of optically-sensed Aphid (Homoptera: Aphidae) wingbeat waveforms. *Annals of the Entomological Society*, 95(1):1–8.

Patnaik, J. L., Juliusson, L., and Vogt, R. L. (2007). Environmental predictors of human west nile virus infections, colorado. *Emerging Infectious Diseases*, 13(11):1788–1790.

Reed, S. C., Williams, C. M., and Chadwick, L. E. (1942). Frequency of wing-beat as a character for separating species races and geographic varieties of drosophila. *Genetics*, 27:349–361.

Roubaud, E. (1918). Rhythmes physiologiques et vol spontan chez l'anopheles maculipennis. *C. R. Hebdomadaires des Seances de l'Academie des Science*, 167:967–969.

Taylor, B. (1998). Biological clocks in mosquitoes. *Website: http://antbase.org/ants/africa/personal/crhtml/covercr.htm*.

Walker, K. (2002). A review of control methods for african malaria vector. Technical Report 108, Bureau for Global Health, Washington, DC, USA.

WHO (2010). The world malaria report. Technical report, World Health Organization.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.