Avaliação de Classificadores para o Reconhecimento Automático de Insetos

Vinícius M. A. de Souza[1], **Diego F. Silva**[1],
Pedro R. P. Garcia[1], **Gustavo E. A. P. A. Batista**[1]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
São Carlos – SP – Brazil

{vsouza, diegofsilva, gbatista}@icmc.usp.br

pedrorael@grad.icmc.usp.br

Abstract. *An important application in public health are the intelligent traps able to selectively capture insect species of interest, such as disease vectors, without affecting other beneficial species to the environment. The implementation on such trap requires the development of a sensor able to automatically detect the species of the insects that enter the trap. Recently, we proposed a sensor that uses lasers and machine learning techniques to automatically classify insects. Aiming to guide the choice of classifiers to be embedded in the intelligent trap, this work presents an experimental evaluation of several classifiers to automatically recognize insect species using the sensor data. Our results indicate that a simple kNN classifier outperforms the proposed baseline and is competitive with more sophisticated techniques such as SVM and GMM classifiers.*

Resumo. *Uma importante aplicação de utilidade pública é o uso de armadilhas inteligentes capazes de capturar seletivamente espécies de insetos de interesse, como as vetores de doenças, sem que espécies benéficas para o meio ambiente sejam afetadas. Para que tal armadilha seja possível, é necessário desenvolver um sensor capaz de detectar automaticamente as espécies dos insetos que entram na armadilha. Recentemente, foi proposto um sensor que utiliza luz laser e técnicas de aprendizado de máquina para classificar insetos automaticamente. Com o objetivo de guiar a escolha dos classificadores a serem embarcados nessa armadilha, este trabalho apresenta a avaliação experimental de diversos classificadores para o reconhecimento automático de insetos a partir de dados coletados pelo sensor. Resultados apontam que um simples algoritmo kNN supera o baseline proposto e é competitivo com técnicas mais sofisticadas como SVM e GMM.*

1. Introdução

A humanidade possui uma estreita relação com insetos, tanto positiva quanto negativamente. De maneira positiva, os insetos possuem um importante papel para o equilíbrio ecológico. Por exemplo, são fontes de alimentos para outras espécies animais, auxiliam na reprodução de espécies vegetais e na produção agrícola ao realizarem o processo de

polinização. Estima-se que os insetos sejam responsáveis pela polinização de aproximadamente 80% das espécies empregadas na agricultura [Dixon 2009]. Além disso, várias espécies têm sido utilizados como bioindicadores de qualidade ambiental, de modo que a sua presença/ausência, distribuição e densidade, permitem definir a qualidade do ecossistema, especialmente em relação a contaminantes do ar, solo e água [Kevan 1999].

Por outro lado, insetos são vetores de doenças que matam milhões de pessoas por ano e deixam outras dezenas de milhões adoecidas. Estima-se que a dengue, doença transmitida por mosquitos do gênero *Aedes*, afete entre 50 e 100 milhões de pessoas por ano e é considerada uma doença endêmica em mais de 100 países [WHO 2009]. A malária, transmitida por mosquitos do gênero *Anopheles*, afeta por volta de 6% da população mundial e estima-se que existam mais de 350 milhões de casos da doença por ano e aproximadamente 2 milhões de casos letais na última década [WHO 2011]. Na agricultura e pecuária, insetos causam enormes prejuízos ao atacar colheitas e animais. Frequentemente estão na raiz do problema chamado insegurança alimentar, referente ao risco de perdas significativas na produção de alimentos [Vreysen and Robinson 2011].

Devido a essa relação de dualidade entre benefícios e malefícios, o desenvolvimento de armadilhas inteligentes é de grande interesse tanto econômico quanto social. Tais armadilhas devem capturar seletivamente determinadas espécies, como as transmissoras de doenças, sem que espécies benéficas sejam afetadas, evitando causar impacto negativo para o meio ambiente. Porém, para capturar insetos seletivamente é necessário que seja realizada a detecção automática das espécies de insetos que entram na armadilha.

Neste sentido, foi proposto recentemente um sensor laser capaz de capturar informações de insetos à distância [Batista et al. 2011]. Este sensor consiste em um feixe de luz laser e um hardware de baixo custo que, aliados a técnicas de mineração de dados, permitem que o sensor identifique automaticamente a espécie de inseto que cruzar o laser. Desde a proposta do sensor, as técnicas de aprendizado de máquina utilizadas para a classificação dos sinais foram pouco exploradas. Desse modo, este trabalho apresenta a avaliação de diferentes classificadores para o reconhecimento automático de insetos. A importância deste estudo está relacionada aos futuros direcionamentos no desenvolvimento do sensor, dado que os melhores resultados obtidos na avaliação apresentada terão suas técnicas adaptadas para que possam ser embarcadas no hardware do sensor.

Este trabalho está organizado da seguinte maneira: na Seção 2 é apresentado o sensor laser utilizado na aplicação de identificação automática de insetos; na Seção 3 são apresentadas as técnicas de aprendizado de máquina utilizadas na avaliação experimental deste trabalho, bem como os atributos extraídos dos sinais obtidos pelo sensor; na Seção 4 são apresentados e discutidos os resultados alcançados; e por fim, na Seção 5 são apresentadas as principais conclusões do trabalho.

2. Sensor Laser para a Identificação de Insetos

Os dados utilizados na avaliação dos algoritmos de classificação deste trabalho foram coletados por um sensor de baixo custo capaz de capturar informações de insetos à distância. Este sensor utiliza componentes como lasers que podem ser facilmente encontrados em lojas de variedades eletrônicas e fototransistores encontrados em controles remotos utilizados em televisores. O esquema ilustrativo do sensor é apresentado na Figura 1-*a*.

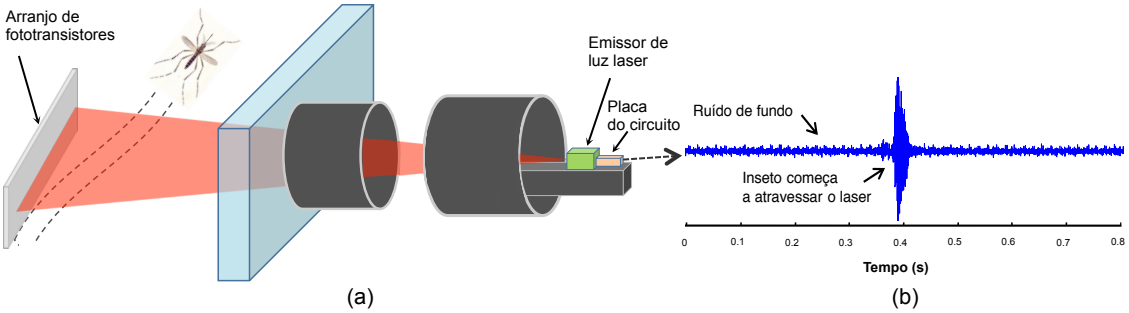

Figura 1. *a*) Esquema ilustrativo do sensor. Uma luz laser planar é direcionada a um arranjo de fototransistores. *b*) Quando um inseto alado cruza a luz, a variação é registrada pelos fototransistores na forma de uma série temporal

O sensor consiste basicamente de um emissor de luz laser planar como os utilizados em serras de madeira para marcar o local de corte, um arranjo de fototransistores e um circuito especialmente projetado para filtrar e amplificar os sinais capturados pelos fototransistores. Quando um inseto alado cruza o laser, o movimento de suas asas ocluem parcialmente a luz causando pequenas variações que são capturadas pelos fototransistores. Essas variações ocorrem em um espaço de tempo e possuem diferentes magnitudes. Assim, a passagem é representada na forma de uma série temporal e, mais especificamente, armazenada como um arquivo de áudio por um gravador digital.

Como pode ser observado na Figura 1-*b*, os dados coletados pelo sensor são constituídos, em geral, de ruído de fundo com "eventos" ocasionais, resultado dos breves cruzamentos do inseto pelo laser. Assim, contar o número de insetos que cruzam a luz pode ser considerada uma tarefa relativamente simples dada a alta razão sinal-ruído, enquanto classificar os sinais de acordo com as possíveis espécies é uma tarefa mais complexa.

2.1. Coleta e Pré-processamento dos Dados

O processo de coleta dos dados foi realizado com sensores acoplados em insetários, de modo que cada insetário contenha insetos de uma única espécie. Dessa maneira, os exemplos são naturalmente rotulados em suas espécies. A coleta foi realizada em condições controladas de laboratório por um período de seis dias. A temperatura e a umidade relativa do ar variaram entre 20°C e 22°C e 20% e 35%, respectivamente.

Após a coleta dos dados é necessária uma etapa de pré-processamento para segmentação e filtragem. A segmentação consiste no uso de um detector de passagens por todo o sinal capturado pelo sensor. O detector utiliza uma janela deslizante sobre os dados e calcula as magnitudes dos componentes do sinal dentro da janela. Então, é utilizada a magnitude máxima dentro de uma faixa de frequências entre 100Hz e 1000Hz (faixa de valores de frequências típicas de batida de asas de insetos) como um valor de confiança para o detector. Dessa maneira, quanto maior for a magnitude, maior a confiança de que o segmento do 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. O funcionamento do detector é ilustrado na Figura 2.

Observa-se na Figura 2 que o detector utiliza uma janela deslizante (W) que corre sobre o sinal capturado pelo sensor. A cada passo, o sinal dentro da janela W é convertido do domínio temporal para o domínio das frequências por meio da Transformada Discreta

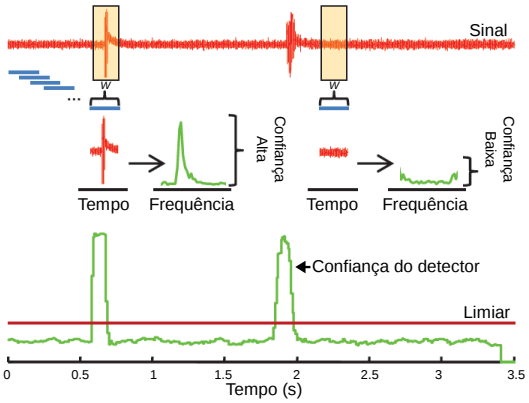

Figura 2. Esquema de funcionamento do detector de eventos [Batista et al. 2011]

de Fourier e extrai-se sua magnitude máxima. Definido o limiar, é possível estimar o nível de confiança de determinada janela conter a passagem de um inseto.

O detector gera fragmentos de áudio com frações de segundo de duração e ao menos uma passagem de inseto, podendo haver mais de uma passagem em um único fragmento. Devido às características físicas do laser, as passagens são acompanhadas de ruído de fundo. A remoção do ruído é importante para que sejam extraídas somente informações relativas às passagens dos insetos, descartando qualquer outro tipo de informação. Para a remoção do ruído, aplicou-se um filtro responsável por remover determinadas faixas de frequência do sinal. Para determinar a faixa de frequência a ser removida, extraiu-se a informação de uma pequena amostra de ruído. Na Figura 3 é exibido um exemplo de fragmento de áudio gerado pelo detector após a etapa de filtragem, obtido pela passagem de um mosquito da espécie *Aedes aegypti*, vetor de doenças como dengue e febre amarela.

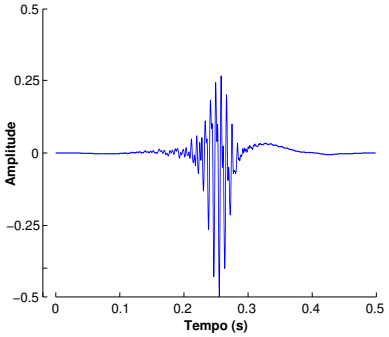

Figura 3. Exemplo de dado filtrado obtido pelo sensor dada a passagem da espécie *Aedes aegypti*

2.2. Descrição do Conjunto de Dados

Após as etapas de coleta e pré-processamento, foram obtidas 5.325 observações para a formação do conjunto de dados avaliado neste trabalho. Cinco diferentes espécies de insetos foram contempladas: as moscas *Drosophila melanogaster* e *Musca domestica* e os mosquitos *Culex quinquefasciatus, Culex tarsalis* e *Aedes aegypti*, principais vetores de doenças como filariose, febre do Nilo Ocidental, dengue e febre amarela. A distribuição das classes é apresentada na Tabela 1.

Tabela 1. Distribuição das classes que constituem o conjunto de dados avaliado

Espécie	Exemplos	Distribuição (%)
Musca domestica	917	7,22
Culex quinquefasciatus	1285	24,13
Culex tarsalis	1265	23,76
Drosophila melanogaster	954	17,91
Aedes aegypti	904	16,98
Total	**5.325**	100

3. Classificação Automática de Insetos

Conforme descrito na Seção 2, os dados gerados pelo sensor laser são arquivos de áudio que podem ser processados como séries temporais. De modo geral, duas abordagens podem ser utilizadas para a classificação de séries temporais: a partir do uso de uma medida de distância que mede a similaridade entre duas séries ou a partir da extração de características intrínsecas às séries que permitam diferenciá-las. Na Subseção 3.1 será discutida a primeira abordagem e a abordagem de classificação a partir da extração de características será introduzida na Subseção 3.2.

3.1. Classificação por Similaridade

A abordagem de classificação de séries temporais por similaridade é simples e efetiva, sendo competitiva com métodos de classificação mais complexos [Ding et al. 2008]. Utiliza-se nesta abordagem uma função de distância $D(T, Q)$ entre duas séries temporais T e Q para encontrar no conjunto de treinamento as k instâncias $T_1, T_2, \ldots, T_k$ mais similares à instância de consulta Q. Assim, a classe predita para Q é a classe mais frequente entre as k instâncias mais similares.

Uma medida frequentemente utilizada para a classificação de séries temporais é distância Euclidiana, definida pela Equação 1.

$$D_{euclidiana}(T, Q) = \sqrt{\sum_{t=1}^{n}(T(t) - Q(t))^2} \quad (1)$$

onde t é o momento em que o fenômeno que gerou a série temporal foi medido e n é o número de observações analisadas.

Originalmente, as séries temporais obtidas pelo sensor laser estão no domínio do tempo (Figura 3), o que impõe restrições para a utilização da distância Euclidiana na busca por similaridade. Por exemplo, as passagens identificadas pelo detector possuem diferentes durações de tempo e conforme pode ser observado na Equação 1, uma restrição da distância Euclidiana é que as séries comparadas devem apresentar o mesmo número de observações. Além disso, a presença de ruído nos dados, mesmo que pequena, é responsável por empobrecer ainda mais a representação. Outro fato a se observar é que há diversas características dos dados que não são evidentes no domínio do tempo, sendo necessária uma representação mais rica que evidencie algum padrão relevante. Esses argumentos foram demonstrados anteriormente em experimentos preliminares utilizando uma versão anterior do sensor [Silva et al. 2011].

Um modo conveniente de superar tais restrições é a utilização da representação no domínio das frequências, denominada espectro de frequências. Essa representação

evidencia as principais componentes de frequência que geraram o sinal e que não são facilmente observáveis no domínio do tempo. Um exemplo da represetanção espectral é ilustrado na Figura 4, dada a passagem de um inseto da espécie *Aedes aegypti*. Na figura é possível observar, por exemplo, a frequência de batida de asas do inseto analisado (646Hz), diretamente relacionada à chamada frequência fundamental do sinal. Outras informações relevantes podem ser encontradas em frequências múltiplas da frequência fundamental, denominadas componentes harmônicas. Destacamos que pequenas diferenças anatômicas entre os insetos podem causar variações nas principais frequências que formam o sinal.

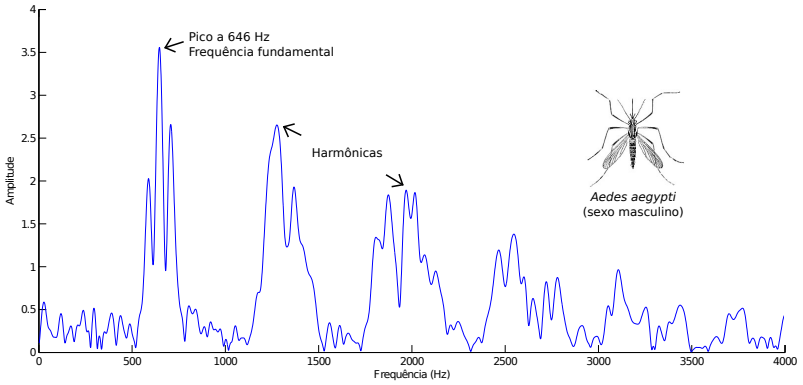

Figura 4. Espectro de frequências dada a passagem de um inseto da espécie *Aedes aegypti* pelo sensor

Dada a importância da localização das componentes de frequência de maior magnitude, é intuitivo pensar que um simples algoritmo com a regra do vizinho mais próximo considerando a similaridade dos espectros de frequência de diferentes sinais pode resultar em um classificador com boa acurácia.

3.2. Classificação por Características Intrínsecas dos Dados

Uma das principais informações presentes nos dados que pode ser utilizada para classificar as passagens em espécies é a frequência de batida de asas dos insetos. Nas últimas décadas, pesquisadores em Entomologia têm medido e analisado como a frequência de batida de asas varia entre as espécies [Oertli 1989, Hyatt and Maughan 1994, Buchwald and Dudley 2010].

A Figura 5 ilustra a variação aproximada para sete espécies de insetos, incluindo três espécies de mosquitos. Nela, é possível observar a ocorrência de sobreposição nas faixas de frequência de batida de asas entre várias espécies. Desse modo, somente essa informação não é suficiente para a caracterização das espécies, sendo necessária a extração de outros atributos intrínsecos às séries temporais que ajudem a distinguir espécies de insetos que possuem valores próximos de frequência de batida de asas.

Existe na literatura uma quantidade imensurável de métodos de extração de atributos que podem ser aplicados nos dados obtidos pelo sensor identificador de insetos. Por isso, foram experimentalmente avaliados diferentes atributos utilizados em aplicações similares que envolvem áudio de curta duração, como o reconhecimento de dígitos isolados [Silva et al. 2012, Silva et al. 2013], reconhecimento de instrumentos musicas

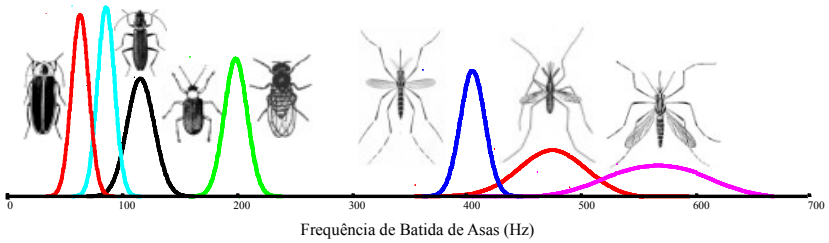

Figura 5. Curvas gaussianas representando média e desvio padrão das frequências de batida de asas de sete espécies de insetos (somente fêmeas). Da esquerda para a direita: *Lucidota atra, Chauliognathus marginatus, Oulema melanopus, Drosophila melanogaster, Culex quinquefasciatus, Anopheles stephensi* e *Aedes aegypti* [Batista et al. 2011]

[Terasawa et al. 2005] e o reconhecimentos de espécies animais por seus cantos ou chamados [Lopes et al. 2011]. Devido às restrições de espaço, serão apresentados neste trabalho somente os resultados obtidos com a extração de um conjunto de atributos. Especificamente, os coeficientes mel cepstrais (*Mel-Frequency Coefficientes* – MFCC), por terem apresentado resultados satisfatórios em nossos experimentos.

Os coeficientes mel cepstrais são extraídos da representação cepstral do sinal utilizando a escala mel. A representação cepstral é obtida por meio da Transformada de Fourier no espectro de frequências do sinal. A escala mel é uma escala que relaciona as frequências físicas às frequências percebidas pelo sistema auditivo humano, introduzida por [Stevens et al. 1937]. A escala mel é baseada no fato de a escala em Hertz não refletir a percepção humana. Por exemplo, um sinal senoidal de 880Hz não soa para o ouvido humano duas vezes mais agudo que um sinal de 440Hz e nem quatro vezes mais agudo que um sinal de 220Hz. A Equação 2 define a conversão da escala de frequências (f) em mel.

$$mel(f) = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

Esse conjunto de atributos é utilizado como entrada para algoritmos de classificação supervisionada como Máquinas de Vetores de Suporte ou Modelo de Misturas Gaussianas, entre outros algoritmos avaliados neste trabalho. Entretanto, a maior parte destes algoritmos possui determinados parâmetros nos quais a otimização de seus valores pode exercer grande influência nos resultados. Por isso, antes da avaliação foi realizada uma etapa preliminar de variação dos principais parâmetros de cada algoritmo avaliado por meio do método de busca em grade (*grid search*) em conjunto com a validação cruzada, conforme discutido em [Hsu et al. 2003].

A busca em grade é um método heurístico que percorre todo o espaço de parâmetros definido pelo usuário. Dados valores de mínimo, máximo e tamanho do passo para cada parâmetro que se deseja otimizar, a influência dessa variação é testada de acordo com alguma medida de avaliação (por exemplo, acurácia) em etapas de validação cruzada. Para reduzir o custo do processo, realiza-se inicialmente uma busca global pelo espaço de parâmetros utilizando-se validação cruzada com um número pequeno de *folds*. Após

encontrar a melhor região do espaço, a busca é refinada nesse subespaço avaliando-se a classificação com a utilização de validação cruzada com 10 *folds*. A configuração de parâmetros que apresenta melhor desempenho na segunda etapa da busca é considerada um ótimo local no espaço de busca definido. Um exemplo ilustrativo é apresentado na Figura 6, em que é considerada a busca pelos valores dos parâmetros A e B. Na imagem, a grade representa o espaço de busca e as cores mais escuras representam as regiões onde a busca é mais refinada.

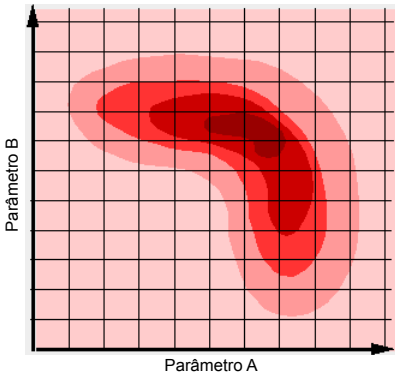

Figura 6. Exemplo de espaço de busca explorado pelo método *grid search*

4. Resultados Experimentais

Nesta seção são apresentados os resultados experimentais obtidos com os dados provenientes do sensor laser identificador de insetos. Os experimentos contemplaram a variação na quantidade de coeficientes mel cepstrais utilizados na avaliação dos algoritmos (Subseção 4.1), a definição de um *baseline* a partir de experimentos de classificação por similaridade e a comparação com diferentes algoritmos de classificação dada a otimização de seus parâmetros (Subseção 4.2).

4.1. Variação na Quantidade de Coeficientes Mel Cepstrais

Comumente, 13 coeficientes mel-cepstrais são aceitos como bom descritor para a classificação de sinais de áudio [Terasawa et al. 2005]. Entretanto, uma quantidade de coeficientes diferente do padrão pode exercer influência positiva nos resultados medidos, como apontado em [Silva et al. 2013]. Desse modo, com o objetivo de encontrar uma quantidade de coeficientes mel cepstrais que seja representativa para os dados avaliados neste trabalho, foi realizado um experimento considerando dois algoritmos de classificação e a variação no número de coeficientes. Os algoritmos utilizados neste experimento foram k-Vizinhos Mais Próximos – kNN (com $k = 1$) e o algoritmo Máquina de Vetores de Suporte – SVM (com $kernel$ RBF, $c = 1$ e $\gamma = 0,01$). A escolha dos algoritmos se deve a resultados satisfatórios obtidos em avaliações preliminares considerando valores padrões para os parâmetros. A quantidade de coeficientes foi variada de 3 a 69 (considerando-se somente valores ímpares). Os resultados obtidos neste experimento podem ser vistos na Figura 7, em que o eixo horizontal denota a quantidade de coeficientes considerada e o eixo vertical corresponde à acurácia obtida por cada configuração.

Nota-se em ambos os algoritmos que a curva que representa a acurácia se mantém crescente para valores superiores a 13 (quantidade tipicamente utilizada) até se estabilizar

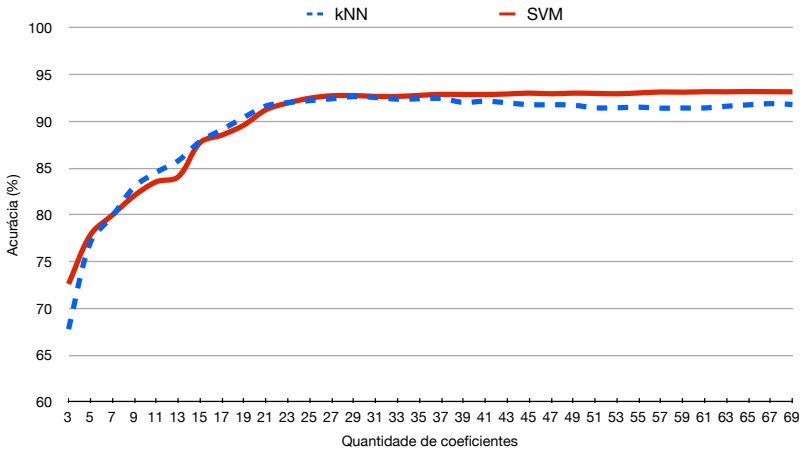

Figura 7. Resultados obtidos pelos algoritmos kNN e SVM dada a variação na quantidade de coeficientes mel cepstrais

ao atingir aproximadamente 25 coeficientes. Embora não seja possível definir pontualmente o melhor número de coeficientes para ambos os algoritmos, é possível notar que a partir de 25 coeficientes as variações são pouco significativas. Além disso, a escolha por uma menor quantidade de coeficientes reduz o tempo gasto nas etapas de extração dos atributos e classificação. Assim, os experimentos deste trabalho que consideram a extração de atributos dos dados consideram um conjunto de 25 coeficientes mel cepstrais para os algoritmos de aprendizado avaliados.

4.2. Avaliação de Algoritmos de Classificação

Conforme anteriomente discutido na Subseção 3.1, a localização das componentes de frequência do sinal fornecem importantes informações sobre a espécie que causou as variações observadas pelo sensor. Por isso, é possível construir um algoritmo baseado na regra do vizinho mais próximo utilizando a similaridade da representação espectral para definir a vizinhança de cada exemplo. O resultado obtido com essa estratégia foi utilizado como *baseline* para os experimentos posteriores. Nessa etapa, os espectros de frequência foram comparados utilizando a distância Euclidiana, descrita na Equação 1. O método *baseline* obteve acurácia de 89,61%. Dada a simplicidade deste método, é esperado que técnicas de classificação mais elaboradas obtenham melhores resultados.

Os resultados de todos os experimentos foram estimados utilizando a técnica de validação cruzada com 10 *folds* em 10 repetições (10x10-*fold cross-validation*). Assim, o processo de validação-cruzada é realizado dez vezes, com diferentes distribuições dos exemplos entre os *folds* a cada execução.

Na Tabela 2 são exibidos os algoritmos de classificação utilizados nos experimentos, bem como os parâmetros variados em cada algoritmo, o intervalo de variação (*início* e *fim*) e o tamanho do passo a ser considerado pelo método de busca em grade. Por exemplo, o algoritmo k-Vizinhos Mais Próximos teve a variação do parâmetro k que representa a quantidade de vizinhos, com *início* $= 1$, *fim* $= 49$ e *passo* $= 2$. Assim, na busca foram considerados todos os valores de $k \in \{1, 3, 5, \ldots, 49\}$. O algoritmo Máquina de Vetores de Suporte teve, além da variação da função de *kernel* (Linear, Polinomial, RBF e Sigmoidal), a variação de mais de um parâmetro em alguns casos. O algoritmo Modelo

de Misturas Gaussianas teve variações na quantidade de gaussianas considerando as covariâncias Esférica, Tied, Diagonal e Completa. Os algoritmos Naive Bayes e Regressão Logística não possuem parâmetros passíveis de variação.

Tabela 2. Algoritmos de classificação utilizados nos experimentos e parâmetros considerados para variação pelo método de busca em grade

Algoritmo	Parâmetros	Variação (início:fim:passo)
k-Vizinhos Mais Próximos	Qtde. vizinhos (k)	k = 1:49:2
Floresta Aleatória	Qtde. árvores (T)	T = 1:100:10
CART (com poda)	Min. instâncias por folha (SL)	SL = 2:10:2
Máquina de Vetores de Suporte ($kernel$ Linear)	Regularização (c)	c = 0,1:2,1:0,5
Máquina de Vetores de Suporte ($kernel$ Polinomial)	Regularização (c); Grau (d)	c = 0,1:2,1:0,5; d =1:10:1
Máquina de Vetores de Suporte ($kernel$ RBF)	Regularização (c); Gamma (γ)	c = 0,1:2,1:0,5; γ = 0,01:0,1:0,01
Máquina de Vetores de Suporte ($kernel$ Sigmoidal)	Regularização (c); Gamma (γ)	c = 0,1:2,1:0,5; γ = 0,01:0,1:0,01
Modelo de Misturas Gaussianas (covariância Esférica)	Qtde. gaussianas (G)	G = 1:10:1
Modelo de Misturas Gaussianas (covariância *Tied*)	Qtde. gaussianas (G)	G = 1:10:1
Modelo de Misturas Gaussianas (covariância Diagonal)	Qtde. gaussianas (G)	G = 1:10:1
Modelo de Misturas Gaussianas (covariância Completa)	Qtde. gaussianas (G)	G = 1:10:1
Regressão Logística	–	–
Naive Bayes	–	–

Com o objetivo de comprovar se os classificadores treinados a partir de coeficientes mel cepstrais são capazes de superar significativamente a simples abordagem utilizada como *baseline*, os resultados foram submetidos a um teste de hipóteses. A primeira etapa consistiu no teste de normalidade dos resultados, ou seja, a verificação se os resultados obtidos podem ser aproximados por uma distribuição normal. Como o teste obteve resultado positivo para todos os classificadores, foi utilizado o teste paramétrico e não pareado *One-way ANOVA* com pós-teste de $Dunnet$. O grau de confiança do teste foi fixado em 95%.

Na Tabela 3 são apresentados os parâmetros obtidos pelo método de busca em grade para cada classificador, bem como os resultados em termos da medida de avaliação acurácia e o resultado do teste de hipótese.

Tabela 3. Resultados da classificação utilizando-se os parâmetros previamente encontrados. Os resultados assinalados com • e ◦ correspondem àqueles que apresentaram melhora ou piora, respectivamente, em relação ao *baseline*, de acordo com o teste de hipótese aplicado

Algoritmo	Parâmetros encontrados	Acurácia
k-Vizinhos Mais Próximos	k = 20	92,14% •
Floresta Aleatória	T = 30	88,37%
CART (com poda)	$SL = 4$	81,18% ◦
Máquina de Vetores de Suporte ($kernel$ Linear)	$c = 0,5$	71,95% ◦
Máquina de Vetores de Suporte ($kernel$ Polinomial)	$c = 0,5$; $d = 7$	90,74%
Máquina de Vetores de Suporte ($kernel$ RBF)	$c = 0,4$; $\gamma = 0,01$	93,51% •
Máquina de Vetores de Suporte ($kernel$ Sigmoidal)	$c = 0,5$; $\gamma = 0,07$	71,77% ◦
Modelo de Misturas Gaussianas (covariância Esférica)	G = 10	92,06% •
Modelo de Misturas Gaussianas (covariância *Tied*)	G = 10	89,32%
Modelo de Misturas Gaussianas (covariância Diagonal)	G = 7	87,19% ◦
Modelo de Misturas Gaussianas (covariância Completa)	G = 8	88,01%
Regressão Logística	–	80,62% ◦
Naive Bayes	–	87,44% ◦
Baseline		***89,61%***

É possível observar na Tabela 3 que dado o resultado de 89,61% de acurácia obtido pelo método *baseline*, somente três de um total de 13 configurações de algoritmos avaliados obtiveram resultados estatisticamente superiores: Máquina de Vetores de Suporte

com o $kernel$ RBF (93,51%), k-Vizinhos Mais Próximos (92,14%) e Modelo de Misturas Gaussianas aplicando covariância Esférica (92,06%). Por outro lado, seis algoritmos apresentaram desempenho estatisticamente inferior, enquanto outros quatro algoritmos apresentaram desempenho similar. A partir destes resultados é interessante notar a influência dos parâmetros em determinados algoritmos. Por exemplo, o algoritmo Máquina de Vetores de Suporte com o *kernel* RBF apresenta resultado estatisticamente superior. Entretanto, o mesmo algoritmo com os *kernels* Linear ou Sigmoidal apresenta resultado estatisticamente inferior. O mesmo comportamento é observado com o algoritmo Modelo de Misturas Gaussianas com covariância Esférica e Diagonal. Outra importante observação é que o algoritmo simples baseado em instâncias k-Vizinhos Mais Próximos apresenta resultado comprovadamente superior ao *baseline* e se mostra competitivo com algoritmos que utilizam técnicas mais sofisticadas como Máquina de Vetores de Suporte e Modelo de Misturas Gaussianas.

5. Conclusões

Este trabalho apresentou a avaliação de classificadores para o reconhecimento automático de insetos. A importância deste estudo está relacionada aos futuros direcionamentos no desenvolvimento de uma aplicação de utilidade pública: armadilhas inteligentes capazes de capturar seletivamente espécies de insetos, como as transmissoras de doenças ou pragas agrícolas.

Diferentes algoritmos de aprendizado de máquina foram experimentalmente avaliados e seus resultados comparados com um simples, porém efetivo, *baseline* obtido a partir de um experimento com a similaridade da representação espectral dos dados. De um total de 13 configurações avaliadas, somente três superaram com diferença estatisticamente significante o resultado obtido pelo método *baseline*.

Com base na avaliação experimental, é possível notar importância da escolha dos parâmetros no desepempenho de algoritmos de classificação mais sofisticados como Máquina de Vetores de Suporte e Modelo de Misturas Gaussianas. Os resultados também apontam que o simples algoritmo baseado em instâncias k-Vizinhos Mais Próximos é competitivo com algoritmos de classificação mais sofisticados e que possuem uma quantidade maior de parâmetros.

Os resultados apresentados neste trabalho devem guiar trabalhos futuros, que incluem a fusão de classificadores, bem como a criação de novos algoritmos mais eficientes em termos de memória e espaço.

Agradecimentos

Os autores agradecem a FAPESP (Processos 2011/17698-5, 2011/04054-2 e 2012/50714-7) e CNPq pelo apoio financeiro concedido.

Referências

Batista, G. E. A. P. A., Keogh, E. J., Mafra-Neto, A., and Rowton, E. (2011). Sensors and software to allow computational entomology, an emerging application of data mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 761–764.

Buchwald, R. and Dudley, R. (2010). Limits to vertical force and power production in bumblebees (hymenoptera: Bombus impatiens). *Journal of Experimental Biology*, 213(3):426–432.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. In *Proceedings of the VLDB Endowmen*, volume 1, pages 1542–1552.

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

Hsu, C. W., Chang, C. C., and Lin, C. J. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.

Hyatt, C. and Maughan, D. (1994). Fourier analysis of wing beat signals: assessing the effects of genetic alterations of flight muscle structure in diptera. *Biophysical journal*, 67(3):1149–1154.

Kevan, P. (1999). Pollinators as bioindicators of the state of the environment: species, activity and diversity. *Agriculture, Ecosystems & Environment*, 74(1-3):373–393.

Lopes, M., Gioppo, L., Higushi, T., Kaestner, C., Silla, C., and Koerich, A. (2011). Automatic bird species identification for large number of species. In *Proceedings of the IEEE International Symposium on Multimedia*, pages 117–122.

Oertli, J. (1989). Relationship of wing beat frequency and temperature during take-off flight in temperate-zone beetles. *Journal of experimental biology*, 145(1):321–338.

Silva, D. F., Batista, G. E., Keogh, E., and Mafra-Neto, A. (2011). Resultados preliminares na classificação de insetos utilizando sensores ópticos. In *Proceedings of the XXXI Congress of the Brazilian Computer Society*, pages 749–760.

Silva, D. F., Souza, V. M. A., and Batista, G. E. A. P. A. (2013). A comparative study between mfcc and lsf coefficients in automatic recognition of isolated digits pronounced in portuguese and english. *(In Press) Acta Scientiarum. Technology*.

Silva, D. F., Souza, V. M. A., Batista, G. E. A. P. A., and Giusti, R. (2012). Spoken digit recognition in portuguese using line spectral frequencies. In *Proceedings of the 13th Ibero-American Conference on Artificial Inteligence*, pages 241–250.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190.

Terasawa, H., Slaney, M., and Berger, J. (2005). The thirteen colors of timbre. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 323–326.

Vreysen, M. and Robinson, A. (2011). Ionising radiation and area-wide management of insect pests to promote sustainable agriculture. a review. *Agronomy for sustainable development*, 31(1):233–250.

WHO (2009). Dengue: guidelines for diagnosis, treatment, prevention and control. Technical report, World Health Organization.

WHO (2011). The world malaria report. Technical report, World Health Organization.