CLASSIFICAÇÃO AUTOMÁTICA EM PASTAGEM USANDO MACHINE LEARNING

ADJ I TCHE [1*], Ibukun Chife Didier, BALICO[2*], Leandro, BENDAHAN [3*], Amaury Burlamaqui

[1]*Graduando na Universidade Federal de Roraima. (Ibukundidier@gmail.com)*

[2] *Professo Ciência da Computação Universidade federal de Roraima.*

[3] *Pesquisador Embrapa Roraima, Amaury.bendahan@embrapa.br*

Palavras Chave: *Pastagem, Insetos-Pragas, Aprendizagem de Máquina, Classificação de imagens.*

INTRODUÇÃO

Na região Amazônica, como em certos países, a produção agropecuária apresenta um grande impacto econômico (HOMMA, 2017). Vários recursos financeiros e operacionais são, por consequência, implementados para assegurar a qualidade da produção e enfrentar degradações de cultivos (PICON et al., 2019). Uns dos principais fatores de degradações que levam importante parte da produção forrageiras, são ataque de insetos-pragas (DIAS-FILHO, 2011).

Em 2015, no município de Alto Alegre no estado de Roraima, ocorreu o ataque de percevejos de gênero *Blissus pulchellus*, o que degradou as pastagens de vários produtores, causando prejuízos econômicos. A aparição do inseto *B. pulchellus* foi o primeiro e o único ainda relatado no Brasil (TEIXEIRA, 2016; DA SILVA, 2019). Esse fato despertou o interesse de órgão de pesquisas, nas medidas preventivas contra a proliferação do inseto pelo resto da país.

Um dos passos dessas medidas é a identificação dos insetos, uma tarefa complexa devido à escassez de taxonomistas especializados na região Norte do Brasil. Entretanto, recentes avanços em áreas como aprendizagem de máquina, possibilitam ferramentas de identificação baseadas em imagens digitais, tendo a acurácia assertiva maior que um especialista (VALAN et al., 2019).

Dessa forma, esse trabalho propõe a avaliação de métodos de aprendizagem de máquina na classificação de insetos causadores de degradação das pastagens no estado de Roraima.

MATERIAL E MÉTODOS

A pesquisa envolveu dois insetos considerados pragas de pastagem no estado de Roraima, são eles: *Blissus pulchellus* e a cigarrinha das pastagens. Por falta de dados suficiente para iniciar a pesquisa, coletou-se primeiramente, os dados no formato de imagens, usando duas câmeras e uma lupa eletrônica Leica EZ 4W (Figura 1c). A primeira câmera foi uma máquina profissional de Modelo CANON EOS REBEL TS, e a segunda um celular SAMSUNG de marca SM-G610M. Sendo um ambiente de experimento controlado, foi construída uma caixa de forma cúbica (50x50x50 cm) (Figura 1a) na qual é colocado pelo fundo da caixa uma lâmpada Led Painel Pop (40x40x4 cm) de 30W de luminosidade difusa 82 Lm/W e na tampa da caixa, 4 lâmpadas Spot Led 12W de tamanho (17x17x3.6 cm) de luminosidade difusa de 75 Lm/W, todas de cor branca (Figura 1b).

As imagens foram tiradas com um fundo na cor magenta primária (#EC008C) em todo o interior da caixa (Figura 1b).

Em seguinte, na etapa de pré-processamento, usando o software Photoshop, iniciou-se o corte e recorte de todas as imagens. Focando nos insetos, aplicou-se as seguintes propriedades nas imagens tiradas da câmera e do celular: temperatura: 4550, colorir: -2, exposição +0,40, vibração +25 e Saturação -100. Foi apenas aplicada nas imagens tiradas da lupa de fundo magenta a Saturação -100.

Logo, depois do pré-processamento, realizou-se a segmentação para remover o fundo das imagens, em seguida a conversão ao tom cinza para a remoção de ruídos. Foi usado nesta etapa, a operação lógica binária AND (&) junto as funções de erosão e de dilatação da Biblioteca OpenCV para uma suavização linear.

Passou-se à extração das características de cor, da textura e de forma, usando as bibliotecas OpenCV e MOHATAS que possua os algoritmos: HU MOMENTS para extrair as formas, HARALICK TEXTURE para as Texturas e COLOR HISTOGRAM para a Cor.

(a) (b) (c)

Figura 1: Caixa Cubica projetado e Lupa eletrônica

A classificação abordada na pesquisa, diz respeito aos algoritmos K-nearest neighbors algorithm – KNN Support Vector Clustering –SVC, RDFC (Random Forest Classificação) e o XGBoost (BEN-HUR, 2008; BREIMAN; LAST; RICE, 2003; CHEN; GUESTRIN, 2016; SCHÖLKOPF et al., 2000; SOUCY; MINEAU, 2001). Os dados coletados foram separados em três categorias: dados de Treinamento, de Teste e de Validação. Os dados de Treinamento e de Teste, processados da mesma forma, entraram respectivamente no treinamento dos algoritmos e na avaliação da aprendizagem destes últimos. Os dados de Validação, se referindo a um caso geral e qualquer de imagens de insetos, não entraram no treinamento dos algoritmos. Assim, servindo para avaliar com que precisão os algoritmos treinados podem classificar ou se adequar aos dados desconhecidos. Para os dados de Validação, a metodologia previu três maneiras diferentes de chegada de dados por usuário, são elas: 1) imagens sem pré-processamento obtida diretamente da Lupa, 2) imagens anteriores segmentadas no intuito de diminuir os custos de processamento e focar no alvo, 3) imagens da categoria 2 usando as mesmas características de pré-processamento conforme aos dados de Treinamento e de Teste.

RESULTADOS E DISCUSSÃO

Os dados de Teste na avaliação da aprendizagem apresentaram as seguintes acurácias: 67.61%, 70.95%, 89.20%, 91.77% para os respectivos algoritmos de SVC, KNN, RDFC, XGBoost (Tabelas: 1, 2, 3 e 4). Avaliando o desempenho de cada algoritmo a respeito das imagens de insetos observamos que com o *B. pulchellus* ficou com uma precisão de 0.76 e um recall de 0.82, e a cigarrinha com uma precisão de 0.58 e um recall de 0.50 para o KNN (Tabela 1). No SVC, obteve-se, respectivamente, uma precisão e recall de: 0.68 e 0.98 para o *B. pulchellus* e 0.69 e 0.07 para a Cigarrinha (Tabela 2). No RDFC, houve uma melhoria de 0.92 e 0.92 para a precisão e recall no Blissus, 0.84 e 0.83 para a Cigarrinha (Tabela 3). Os resultados com o algoritmo XGBoost foram superiores com uma precisão e recall de: 0.93 e 0.95 no Blissus, e no 0.90 e 0.85 na Cigarrinha (Tabela 4).

KNN			
Acurácia: 70.95%			
	Precision	Recall	Num
B. pul	0.79	0.82	258
Ciga	0.58	0.50	131

Tabela 1: Resultados-Dados de Teste com KNN.

RDFC			
Acurácia: 89.20%			
	Precision	Recall	Num
B. pul	0.92	0.92	258
Ciga	0.84	0.83	131

Tabela 2: Resultado-Dados de Teste com RDFC.

SVC			
Acurácia: 67.61%			
	Precision	Recall	Num
B. pul	0.68	0.98	258
Ciga	0.69	0.07	131

Tabela 3: Resultados-Dados de Teste com SVC.

XGBoost			
Acurácia: 91.77%			
	Precision	Recall	Num
B. pul	0.93	0.95	258
Ciga	0.90	0.85	131

Tabela 4: Resultado-Dados de Teste com XGBoost.

Assim, essas medidas apontam o XGBoost como o melhor classificador nestas tarefas, logo depois o RDFC e o KNN, pela alta pontuação dos índices de precisão e de recall o que significa, respetivamente, o quanto um algoritmo prever como certa a sua predição e o quanto, realmente, essa predição é verdadeira. No caso de SVC e do KNN os algoritmos conseguiram bem aprender um pouco acima da média, as características do B. pulchellus, medianamente as características da cigarrinha no KNN e pouco da característica da cigarrinha no SVC (precisão: 0.69 e recall: 0.07). Isso leva a descartar o SVC e considerar como aceitável nesta tarefa o KNN.

Os resultados obtidos com a validação do reconhecimento dos insetos utilizando o algoritmo KNN mostraram que sem segmentação houve uma acurácia de 76 % com uma precisão de 0.94 (proporção preditiva positiva) e um recall de 0.7 (proporção de resultados reais positivo) para o *B. pulchellus*. No que diz respeito a cigarrinha, os resultados foram, respectivamente, de 0.10 e 0.32 de precisão e recall (Tabela 5).

Utilizando a segmentação das imagens de validação, a acurácia obtida foi de 79,67%, com uma precisão de 0.89 e recall 0.88 para o Blissus e com a cigarrinha, uma precisão de 0.24 e um recall 0.25 (Tabela 5).

Quando foram usadas a saturação com as imagens segmentadas a acurácia caiu para 60,88 % tendo com uma precisão e recall para o Blissus os seguintes valores: 0.91 e 0.61. Com a cigarrinha a precisão foi de 0.10 e o recall 0.53 (Tabela 5).

KNN									
	SEM SEGMENTAÇÃO			COM SEGMENTAÇÃO			SATURADA E SEGMENTADA		
	Acurácia: 76,06%			Acurácia: 79,67%			Acurácia: 60,83%		
	Precision	Recall	Num	Precision	Recall	Num	Precision	Recall	Num
B. pul	0.94	0.79	265	0.89	0.88	265	0.94	0.61	200
Ciga	0.10	0.32	19	0.24	0.25	40	0.10	0.53	17

Tabela 5: Resultados das três maneiras diferentes de validação do algoritmo KNN no reconhecimento dos insetos Blissus pulchellus e cigarrinha das pastagens.

Destes resultados, observa-se que, apesar das acurácias serem acima da média, o KNN não teve um bom desempenho com os dados de imagem saturadas e segmentadas nem com as imagens sem segmentação. Isso constata-se nas observações dos valores fracos da precisão e do recall. As imagens com segmentação no KNN apresentaram precisão e recall altos para o *B. pulchellus*, o que significa que o algoritmo conseguiu realmente reconhecer o inseto, e, com a cigarrinha, apesar da precisão estar bem próxima ao recall, os dois valores são fracos. Assim, pode-se dizer que o KNN não conseguiu identificar toda a categoria de cigarrinhas.

Quando se utilizou o algoritmo SVC, obteve-se as seguintes acurácias, 93%, 86,89%, 92,17%, respectivamente, para os dados de Validação sem segmentação, com segmentação e com imagens saturadas e segmentadas (Tabela 6). Com as imagens do Blissus, obteve-se uma precisão de 0.93 e um recall de 1.00 para imagens sem segmentação, em seguida, uma precisão de 0.87 e um recall de 1.00 para imagens segmentadas, e obteve-se finalmente, uma precisão de 0.92 e um recall de 1.00 nas imagens saturadas e segmentadas (Tabela 6). Esses resultados apontam uma grande precisão do algoritmo SVC no reconhecimento do *B. pulchellus*. No que diz respeito à precisão e ao recall das imagens de validação da

cigarrinha, o valor 0.00 foi obtido em todas tentativas (Tabela 6). De fato, o algoritmo SVC não consegue reconhecer as imagens de Validação da cigarrinha.

SVC									
	SEM SEGMENTAÇÃO			COM SEGMENTAÇÃO			SATURADA E SEGMENTADA		
	Acurácia: 93,31%			Acurácia: 86,89%			Acurácia: 67,61%		
	Precision	Recall	Num	Precision	Recall	Num	Precision	Recall	Num
B. pul	0.93	1.00	265	0.87	1.00	265	0.92	1.00	200
Ciga	0.00	0.00	19	0.00	0.00	40	0.00	0.00	17

Tabela 6: Resultados das três maneiras diferentes de validação do algoritmo SVC no reconhecimento dos insetos Blissus pulchellus e cigarrinha das pastagens

A classificação das imagens com o algoritmo RDFC deram como acurácias os valores 7.75%, 85,57% e 25.81% para, respectivamente, os dados de validação sem segmentação, com segmentação e com saturação e segmentação (Tabela 7). Isso aponta que o algoritmo RDFC reconhece os dados de Validação com segmentação. Olhando de perto os índices de desempenho, com os dados de validação sem segmentação a precisão e o recall ficaram, respectivamente, com os valores de 0.80 e 0.02 para o Blissus, e de 0.06 e 0.95 para a cigarrinha (Tabela 7). Note-se assim, que o algoritmo reconheceu pouco as características do *B. pulchellus* e ficou muito tendencioso com as características da cigarrinha. Nos dados de Validação com segmentação, a precisão e o recall ficaram, respectivamente, com os valores de 0.87 e 0.97 para o Blissus, e de 0.30 e 0.07 para a cigarrinha (Tabela 7). Isto significa que o algoritmo consegue reconhecer as imagens do *B. pulchellus* quando for segmentada e não consegue reconhecer as da cigarrinha. Quanto às imagens saturadas e segmentadas, a precisão deu 1.00 e o recall 0.20 para o Blissus e para a cigarrinha, a precisão e o recall deram 0.10 e 1.00 (Tabela 7). Assim, observe-se, também, uma orientação tendenciosa a prever os insetos do *B. pulchellus* desta base de validação como cigarrinha pelo RDFC.

RDFC									
	SEM SEGMENTAÇÃO			COM SEGMENTAÇÃO			SATURADA E SEGMENTADA		
	Acurácia: 7,75%			Acurácia: 85,57%			Acurácia: 25,81%		
	Precision	Recall	Num	Precision	Recall	Num	Precision	Recall	Num
B. pul	0.80	0.02	265	0.87	0.97	265	1.00	0.20	200
Ciga	0.06	0.95	19	0.30	0.07	40	0.10	1.00	17

Figura 7: Resultados das três maneiras diferentes de validação do algoritmo RDFC no reconhecimento dos insetos Blissus pulchellus e cigarrinha das pastagens

Com o algoritmo XGBoost, obteve-se, respectivamente, para os dados de Validação sem segmentação, com segmentação e com saturação e segmentação, as acurácias de 72,5%, 86,9% e 32,26% (Tabela 8). Com essas acurácias os dados de Validação com segmentação se posicionam como os mais reconhecidos na tarefa de classificação. No *B. pulchellus*, obteve-se, com as imagens sem segmentação, uma precisão de 0.92 e um recall de 0.78, em seguida, com as imagens com segmentação, uma precisão de 0.87 e um recall de 1.00 e, por fim, com as imagens saturadas e segmentadas, uma precisão de 0.98 e 0.27 (Tabela 8). A precisão e o recall de imagens de validação sem segmentação e com segmentação, apontam que o algoritmo XGBoost se encaixe no reconhecimento do inseto. O fato de obter um recall fraco com os dados de validação saturada e segmentada aponta que o algoritmo está tendencioso a reconhecer as imagens do *B. pulchellus*. Analisando os resultados da cigarrinha, verificou-se uma precisão e um recall praticamente igual a zero, com os dados de validação sem segmentação e com segmentação. Com os dados de validação saturada e segmentada, obteve-se uma precisão muito fraca (0.10) e, ao contrário, um recall alto (0.94) (Tabela 8). Pode-se notar, então, que o algoritmo de XGBoost é ineficaz para o reconhecimento dos insetos destas bases de dados.

XGBOOST									
	SEM SEGMENTAÇÃO			COM SEGMENTAÇÃO			SATURADA E SEGMENTADA		
	Acurácia: 72,54%			Acurácia: 86,89%			Acurácia: 32,26%		
	Precision	Recall	Num	Precision	Recall	Num	Precision	Recall	Num
B. pul	0.92	0.78	265	0.87	1.00	265	0.98	0.27	200
Ciga	0.03	0.00	19	0.00	0.00	40	0.10	0.94	17

Figura 8: Resultados das três maneiras diferentes de validação do algoritmo XGBoost no reconhecimento de dos insetos Blissus pulchellus e cigarrinha das pastagens.

Em geral, os dados de validação foram observados situações de overfitting (precisão alta, recall fraca) e de underfitting (precisão baixa, recall alta) que são, simplesmente, quando o algoritmo tende a reconhecer ou aprender muito ou pouco das características das imagens. No que diz respeito as categorias de dados da validação, pode-se concluir que as imagens apenas segmentadas apresentaram um bom resultado a respeito de suas acurácias e também no seu desempenho, particularmente com o *B. pulchellus*. Em relação à cigarrinha, certos algoritmos tiveram valores de desempenho bem fraco, praticamente igual a zero. Houve carência de informação que contribua na classificação das imagens da cigarrinha. Os algoritmos ficaram bons na sua tarefa de identificação, mas para não ocorrer problemas de overfitting ou de underfitting alguns ajustes podem ser feitos, como por exemplo o aumento da imagem da base de dados.

CONCLUSÕES

Os resultados apontam uma viabilidade de classificação dos insetos-pragas *B. pulchellus* e Cigarrinha pelos algoritmos de aprendizagem de máquinas com as acurácias indo de 67% para 91% com os dados de Teste. O XGBoost e o RDFC se posicionaram como os melhores, tanto pela acurácia e os índices de medida do desempenho. Ainda se apresentaram, ótimos com os dados de Validação que não entraram na aprendizagem apenas com imagem segmentadas.

AGRADECIMENTOS

Agradecimento a Universidade Federal de Roraima e ao departamento de ciência da computação, A Embrapa Roraima, e ao Fundo Amazônia, por financiar parte dessa pesquisa, por meio do Projeto Estratégias para o desenvolvimento sustentável da agricultura familiar com enfoque em sistemas agroflorestais e recuperação de pastagens degradadas na região amazônica (AMAPEC).

BEN-HUR, A. Support vector clustering. **Scholarpedia**, v. 3, n. 6, p. 5187, 2008.

BREIMAN, L.; LAST, M.; RICE, J. **Random Forests: Finding Quasars**. Statistical Challenges in Astronomy. **Anais**... In: DESAFIOS ESTATÍSTICOS EM ASTRONOMIA. Nova York: Springer, 2003Disponível em: <https://link.springer.com/book/10.1007/b97240#about>

CHEN, T.; GUESTRIN, C. **Xgboost: A scalable tree boosting system**. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. **Anais**...2016

DIAS-FILHO, M. **Degradação de pastagens: processos, causas e estratégias de recuperação**. 4. ed. Belém, PA: Do Autor, 2011.

HOMMA, A. K. O. A Terceira Natureza da Amazônia. **Revista Paranaense de Desenvolvimento**, v. 38, n. 132, p. 27–42, 2017.

PICON, A. et al. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. **Computers and Electronics in Agriculture**, v. 161, p. 280–290, 2019.

SCHÖLKOPF, B. et al. New support vector algorithms. **Neural computation**, v. 12, n. 5, p. 1207–1245, 2000.

SOUCY, P.; MINEAU, G. W. **A simple KNN algorithm for text categorization.** Proceedings 2001 IEEE International Conference on Data Mining. **Anais**... In: 2001 IEEE INTERNATIONAL CONFERENCE ON DATA MINING. San Jose, CA, USA: IEEE Comput. Soc, 2001Disponível em: <http://ieeexplore.ieee.org/document/989592/>. Acesso em: 25 set. 2020

VALAN, M. et al. Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. **Systematic biology**, v. 68, n. 6, p. 876–895, 2019.