Definindo e Predizendo Níveis de Saúde de Colônias de Abelhas via Clusterização e Classificação

Antonio Rafael Braga[1], Daniel A. Silva[2], Juvêncio S. Nobre[2], Breno M. Freitas[3], Danielo G. Gomes[1]

[1]Departamento de Engenharia de Teleinformática, Universidade Federal do Ceará (UFC).

[2]Departamento de Estatística e Matemática Aplicada, UFC.

[3]Setor de Abelhas, Departamento de Zootecnia, UFC.

{danielamaral}@alu.ufc.br, {rafaelbraga,juvencio,freitas,danielo}@ufc.br

Resumo. *As abelhas são essenciais à produção de alimentos para o ser humano e para manutenção dos ecossistemas. Esse artigo apresenta uma solução para calcular os níveis de estados de saúde de colônias de abelhas usando dados de sensores internos e externos à colônia e de inspeções in loco realizadas por apicultores. A clusterização foi usada para determinar a quantidade de níveis de saúde e a classificação para criação de um modelo de predição. Obteve-se um modelo de classificação com taxa de acerto de 99.36%.*

Abstract. *Bees are essential for the production of food for humans and the maintenance of ecosystems. This paper presents a solution for calculating bee colony health status levels using data from internal and external colony sensors and on-site inspections by beekeepers. Clustering was used to determine the number of health levels and classification to create a prediction model. We obtained a classification model with an accuracy ratio of 99.36%.*

1. Introdução

As abelhas são essenciais à produção de alimentos para o ser humano e para manutenção dos ecossistemas pois constituem o grupo mais importante de polinizadores [Potts et al. 2016]. No entanto, a diversidade de polinizadores está em declínio. Em paisagens agrícolas, isso é freqüentemente observado em grandes monoculturas. Existe a preocupação de que a expansão urbana e o desmatamento também estejam impactando negativamente a diversidade de polinizadores. Reduções na diversidade e abundância de polinizadores podem afetar a reprodução de espécies de plantas, a produção agrícola, a segurança alimentar e o bem-estar humano [Potts et al. 2010].

Assim, o monitoramento via Rede de Sensores sem Fio tem sido usado mais recentemente para evitar a morte de colônia de abelhas. Associado ao monitoramento, os apicultores fazem uso também de inspeções *in-loco*. Contudo, essas inspeções são prejudiciais à saúde da colônia. Logo, torna-se relevante a predição do nível de saúde de uma colônia para notificação do apicultor (usuário final) para que o mesmo possa atuar através de manejos para restituição da saúde da colônia.

Nesse contexto, este trabalho apresenta uma solução para identificação da quantidade ótima de estados de saúde de colônias de abelhas e dos tipos de estados baseada

em dados de sensores internos e externos às colônias e de inspeção previamente realizadas. A solução utiliza a clusterização (aprendizagem não-supervisionada) e a classificação (aprendizagem supervisionada).

Na etapa de clusterização, o índice de validação *Calinski-Harabasz* (CH) [Calinski and Harabasz 1974] e o algoritmo *k-means* [MacQueen 1967] foram usados para determinar a quantidade ideal de classes de saúde das colônias. Os dados de inspeções *in loco* foram utilizados para fazer a rotulagem dos dados de sensores em relação ao estado de saúde. Após a rotulagem, foram treinados 4 algoritmos de classificação distintos, o k-vizinhos mais próximos (*k-Nearest Neighbors*, k-NN), as Florestas Aleatórias (*Random Forest*, RF), Redes Neurais (*Neural Networks*, NN) e Máquinas de Vetores Suporte (*Support Vector Machine*, SVM).

2. Materiais e Métodos

2.1. O Conjunto de Dados

As colméias utilizadas nesta pesquisa pertencem ao projeto "*Bayer Bee Care*" (https://www.agro.bayer.com.br/beecare) (BBC) e estão localizadas no estado da Carolina do Norte, EUA. Foram usadas 4 colméias, duas do apiário 'BBCC', em Durham, monitoradas no periodo de Junho de 2017 à Abril de 2019, e 2 no apiário Beesboro em Clayton, monitoradas no periodo de Junho de 2017 à Junho de 2018. Ambas possuem abelhas da espécie *Apis mellifera*. Os dados de sensores internos às colméias foram coletados através do sistema Brood Minder (https://broodminder.com/). Os dados de sensores externos foram obtidas de estações meteorológica do Serviço Nacional de Meteorologia dos EUA. Para o apiário BBCC, foi usada a estação do aeroporto internacional de Raleigh-Durham e para o apiário Beesboro a estação do aeroport Johnston County.

Os dados de inspeções foram obtidos com o auxílio do "*Healthy Colony Checklist* (HCC)", que é um documento usado para padronizar a inspeção de uma colônia de abelhas por um apicultor. As inspeções foram realizadas semanalmente. O HCC[1] utilizado neste estudo foi o proposto pelo projeto BBC. O HCC é composto por 6 fatores binários para definir o estado de saúde da colônia, são eles: 1 - Brood (Ninhada), 2 - Bees (Abelhas Adultas), 3 - Queen (Rainha), 4 - Food (Alimento), 5 - Stressors (Estressores) e 6 - Space (Espaço). Assim, se todos os itens forem marcados como "sem problema", então a colônia é considerada 100% saúdável. Por outro lado, cada item marcado como 'com problema', representa uma diminuição de 1/6 do nível de saúde da colônia. Esse nível de saúde calculado é utilizado como variável dependente, classes.

2.2. Pré-processamento

O conjunto de dados foi disponibilizado em arquivos separados por dados de sensores internos, externos e dados de inspeção. Os dados disponibilizados, foram divididos por Apiário e subdivididos por Colméia. A Tabela 1 descreve o conteúdo de cada arquivo disponibilizado.

Assim, para o merge dos arquivos foi utilizada a variável de tempo *Human_timestamp* com um limite de similaridade de 57 minutos para os dados de sensores

[1]https://beehealth.bayer.us/bayer-news-and-resources/setting-the-standard-for-managing-healthy-honey-bee-colonies

Arquivo	Variáveis
Sensores Internos	Human_timestamp, Brood_Temp, Brood_Humidity, Hive_Temp, Hive_Humidity, Scale_Temp, Scale_Humidity, Weight, Hive e Apiary
Sensores Externos	Human_timestamp, Temperature, DewPoint, Pressure, WindDirection, WindSpeed, Sky-Condition, Precipitation1Hr, Precipitation6Hr e Apiary
Inspeção	Human_timestamp, Apiary, Hive, Brood, Bees, Queen, Food, Stressors, Space e Code

Tabela 1. Conteúdo dos dados de sensores internos, externos e de inspeção

para mais ou para menos e para o merge dos dados de inspeção um limite de 1 semana para mais ou para menos. Obtivemos então, um único arquivo com 15.490 amostras contendo as variáveis de sensores internos, externos e de inspeções (rótulo da amostra). Após, removemos algumas colunas não-informativas, as causas de remoção de cada variável são descritas na Tabela 2.

Variáveis	Motivo de Remoção
Hive, Apiary	Queremos um modelo geral, portanto, um modelo que diferencia apenas 2 apiarios e suas colmeias não é o adequado
ScaleTemp, ScaleHumidity	Variáveis sem sentido, já que o sensor Scale só nos dá a informação de peso da colônia.
Pressure	Todos os valores faltantes ocorrem no Apiário "Beesboro". Portanto, a imputação não é uma boa idéia
SkyCondition, Precipitation1Hr, Precipitation6Hr	Alto índice de valores faltantes ($>$90%)

Tabela 2. Motivos de Remoção de variáveis no conjunto de dados

Após a remoção das variáveis não-informativas, notamos diferentes padrões de medição e de escalonamento. Para fins de padronização do experimento transformamos as unidades das variáveis de temperatura de $Fahrenheit$ para $Celsius$ e as variáveis de peso de $Pounds$ para $Kilo$. Após as conversões, reescalonamos as variáveis $Temperature$ e $DewPoint$. Para a consistencia dos dados definimos um limite de -10º à 50º para as variáveis de temperatura, 0 à 100 para as variáveis de humidade e 1 à 100 Kg para a variável peso. Caso a variável em estudo não se enquadre nesses limites ela será imputada pelo método *Multivariate Imputation by Chained Equations* (MICE) [Buuren and Groothuis-Oudshoorn 2010].

Como existem variáveis com medidas diferentes e com escalas e limites muito diferentes, foi realizada a padronização dos dados através da transformação z-score. Para adicionar mais informações aos modelo de predição, foram adicionadas duas variáveis independentes, são elas: Season (Estação do Ano) e TurnDay (Turno do Dia). Por fim, foi feita a conversão das variáveis nominais Season e TurnDay na abordagem Casela de Referência.

3. Algoritmos Propostos

A seguir, serão apresentados os algoritmos propostos nesse trabalho. O Algoritmo 1 nos fornece o melhor valor do número de clusters com base no indice CH, em síntese, esse algoritmo recebe vários candidatos ao número de clusters e para cada candidato o laço for (l. 2-9) agrupa o conjunto de dados de interesse pelo algoritmo k-means (l. 3-7). Para esse candidato em específico é calculado o índice CH associado (l. 8). Então, o melhor valor do número de clusters é aquele com maior valor associado ao vetor de índices CH.

O Algoritmo 2 utiliza o resultado do Algoritmo 1 para realizar uma busca exaustiva do melhor agrupamento, em síntese, esse algoritmo recebe um valor ótimo k do número de cluster, fatores de inspeção e um conjunto de dados. É inserido, então, todos

os possíveis agrupamentos em k clusters dos fatores de inspeção (l. 2). No laço for (l. 5-10) é realizada a rotulagem do conjunto de dados e o cálculo da acurácia de um classificador genérico. A rotulagem (l. 6-7) é realizada através de uma associação direta entre a quantidade de itens de inspeção saúdáveis e a saúde da colônia, ver Seção 2.1. Ao final do algoritmo, é retornada um vetor com o desempenho/acurácia de cada agrupamento de k clusters pelo modelo de classificação escolhido. A partir desses resultados pode se escolher o melhor agrupamento de fatores e o melhor algoritmo a ser utilizado.

Algoritmo 1 Algoritmo para a escolha do melhor valor de k via *k-médias* e o índice CH

Entrada: **K** (vetor com os possíveis números de clusters)
D (um conjunto de dados)
Saída: Um vetor com os índices CH associados ao vetor **K**
1 indices = {}
2 para cada $k \in$ **K faça**
3 **escolha** *k* observações de **D** como os centróides iniciais
4 **repita**
5 **atribua** cada observação de **D** ao cluster com maior similaridade, baseada na média dos objetos no cluster;
6 **atualize** as médias em cada cluster com as observações realocadas;
7 **até convergência**
8 **insira** em **indices** o índice de *CH* associado aos *k* clusters
9 fim
10 retorne indices

Algoritmo 2 Escolhe a melhor configuração de agrupamento dos fatores de inspeção através de um classificador genérico **C**

Entrada: **k** (quantidade de classes desejadas)
F (fatores de inspeção)
D (conjunto de dados correspondente à inspeção)
Saída: Um vetor de acurácias obtidas por um classificador genérico associadas a cada possivel agrupamento de tamanho k
1 acuracias = {}
2 agrupamentos = possiveis agrupamentos de fatores
4 soma = soma dos fatores de inspeção
5 para cada *agrupamento* $\in$ **agrupamentos faça**
6 **classe** = {} (classe ou estado de saúde a ser atribuida)
7 **atribua** o valor da classe (0 à k) a variável **classe**, baseada na soma dos fatores de inspeção e no agrupamento da iteração atual
8 **treine** o *classificador genérico* **C** utilizando o conjunto de dados **D** combinado à **classe**, como preditora, em um experimento de validação cruzada.
9 **insira** em **acuracias** a métrica acurácia correspondente ao experimento de validação cruzada na iteração atual
10 fim
11 retorne acuracias

3.1. Validação Cruzada e Ajuste dos Hiper-parâmetros

Dado que já sabemos o valor ótimo do número de clusters ou classes, o Algoritmo 2 entra em ação. Para cada possível agrupamento, são avaliados os 4 classificadores. Em cada avaliação do classificador $j \in \{$k-NN, RF, NN e SVM$\}$, é feito um experimento de validação cruzada de 10-dobras (10-fold) e avaliado a acurácia com um vetor H_j de hiperparâmetros possiveis. O vetor H_j para cada classificador é definido na Tabela 3,

Classificador	Hiperparâmetros Fixos	Hiperparâmetros possíveis
k-NN	Nenhum	k: {1,2, ..., 30}
Random Forest	Nenhum	mtry: {2, 7, 12}, splitrule: {gini, extratrees}
Neural Network	MLP, dim: (400, 200), dropout: 0.45, batch_size: 100, optimizer: Adam	Nenhum
SVM	kernel: RBF, gamma: 1.5, C = 9	Nenhum

Tabela 3. Configurações dos classificadores: k-NN, RF, NN e SVM

4. Resultados

Para execução do Algoritmo 1, foi utilizado um vetor K para os possíveis números de clusters onde $K = 3, ..., 10$. Essa definição se deu com o objetivo de caracterizar no mínimo à partir de 3 estados de saúde, que podem ser entendidos como: saudável, alerta e doente. Para $k = 2$ obtem-se uma classificação binária, em que as duas possíveis classes são extremas e, portanto, não há uma classe que sirva de alerta para o usuário final. O resultado pode ser observado na Figura 1.

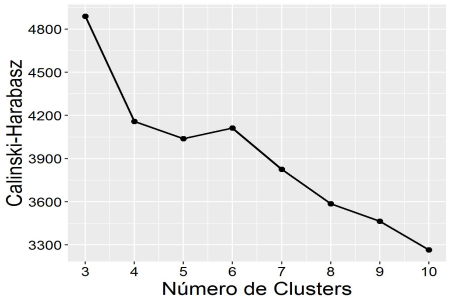

Figura 1. Gráfico resultante da aplicação do Algoritmo 1

O melhor valor de k é 3. Após a definição do numero de classes ($k = 3$), foi possível executar o Algoritmo 2 a fim de definir o melhor agrupamento de fatores de inspeção. O número de possíveis agrupamentos é igual 90. Os resultados são mostrados na Tabela 4, onde a coluna "$Agrupamento$" fornece a identificação de cada possível agrupamento.

Os agrupamentos estão ordenados de acordo com os maiores mínimos de sensibilidade e especificidade interclasses. Essa abordagem nos proporcionará uma configuração de clusters que seja menos penalizada pelos algoritmos no experimento e com alta acurácia. Para escolher o melhor agrupamento tomaremos como base um *trade-off* entre acurácia, sensibilidade e especificidade. Embora, o agrupamento 2456 1 3 nos forneça a maior acurácia dos agrupamentos (99.36%), esse resultado não é o melhor, pois, o que ocasiona essa alta acurácia é o excesso de fatores agrupados. O mesmo ocorre para os agrupamentos 1456 2 3 e 3456 1 2.

Assim, o melhor agrupamento é 2 13 456, que possui o melhor *trade-off* entre acurácia, sensibilidade e especificidade com uma acurácia de 99.23% nas Redes Neurais. É possível observar ainda que as melhores acurácias estão relacionadas à presença dos itens de inspeção 4, 5 e 6 em um mesmo cluster. Assim, esses três itens poderiam ser agrupados para indicação de um nível de saúde. Bem como os itens 1 e 3. Portanto, um alerta poderia ser emitido caso qualquer item do grupo 456 ou 13 apresente um problema.

Vale destacar que os itens 4, 5 e 6 da planilha de inspeção são itens que não estão diretamente ligado à colônia em si, Food (Alimento), Stressors (Estressores) e Space (Espaço), respectivamente. O agrupamento formado pelos itens 1 e 3 possui os itens da planilha de inspeção diretamente ligados ao à rainha. Uma vez que o item 1 (Brood - Ninhada) representa todas as fases da ninhada e ínstares.

5. Conclusões e Trabalhos Futuros

Assim, foi possível obter um modelo de classificação de alta precisão, com taxa de acero de até 99,36%. Foi possível também determinar o melhor agrupamento dos itens da inspeções para definição do nível de saúde de uma colônia de abelhas. Como trabalho futuro, pretende-se aplicar a solução proposta em um conjunto de dados maior.

Agradecimentos

Esse estudo foi financiado em parte pela CAPES - código de financiamento 001. Danielo G. Gomes e Breno M. Freitas agradecem o suporte financeiro do CNPq, processos #302934/2010-3, #311878/2016-4 e #432585/2016-8.

Agrupamento	Acurácia (%)			
	k-NN	R. Forest	NN	SVM
5 12 346	90.11	95.09	91.48	90.73
2 13 456	**98.55**	**99.21**	**99.23**	**98.32**
12 35 46	90.27	94.89	92.55	90.67
2 15 346	89.87	94.89	90.90	90.59
13 25 46	90.18	94.80	91.84	90.21
15 23 46	89.75	94.83	91.61	90.50
1246 3 5	90.50	95.04	91.22	91.12
2 46 135	89.88	94.83	91.61	90.31
13 24 56	94.29	96.69	95.35	94.78
5 23 146	89.57	94.80	91.44	90.21
5 46 123	90.19	95.15	91.97	90.61
2 35 146	89.77	94.88	90.27	90.37
5 13 246	89.95	95.02	92.17	90.85
1456 2 3	**98.50**	**99.11**	**98.60**	**98.49**
2 56 134	94.04	96.65	95.85	94.75
3 56 124	94.56	96.79	96.08	95.07
3 15 246	89.88	94.87	91.54	90.47
1346 2 5	89.97	94.93	92.27	90.71
3 25 146	89.72	94.67	92.40	90.21
2 34 156	94.06	96.65	95.09	94.63
3 24 156	94.26	96.73	95.22	94.76
3 12 456	**98.75**	**99.25**	**99.16**	**98.61**
12 34 56	94.29	96.76	96.30	94.69
3 46 125	90.20	94.87	91.09	90.42
5 34 126	88.88	94.23	90.73	90.01
3 45 126	92.07	95.42	92.96	92.57
1 23 456	**98.52**	**99.12**	**99.16**	**98.33**
14 23 56	93.94	96.56	95.72	94.28
3 14 256	93.95	96.45	95.44	94.54
2 14 356	93.93	96.62	95.50	94.51
4 35 126	88.99	94.13	90.96	89.83
4 23 156	93.91	96.71	95.57	94.25
4 12 356	94.34	96.73	95.65	94.62
2456 1 3	**99.06**	**99.36**	**99.16**	**98.99**
4 13 256	94.00	96.64	95.39	94.38
4 56 123	94.26	96.80	95.80	94.59
1256 3 4	94.28	96.66	95.76	94.86
5 14 236	88.53	94.00	90.40	89.55
4 15 236	88.40	94.09	89.93	89.59
1236 4 5	89.13	94.19	90.85	89.94
4 25 136	88.63	93.98	90.73	89.37
14 26 35	88.66	93.99	90.98	89.50
1 24 356	94.31	96.83	95.44	94.84
2356 1 4	94.15	96.58	95.59	94.60
5 26 134	88.82	94.17	91.18	90.21
1356 2 4	93.82	96.53	95.93	94.49
3 26 145	91.83	95.29	93.80	92.38
5 24 136	88.78	94.22	91.03	89.95
2 45 136	91.89	95.40	93.44	92.26
13 26 45	91.80	95.28	93.37	92.36
3456 1 2	**98.68**	**99.32**	**99.03**	**98.72**
1 34 256	94.04	96.58	95.01	94.74
1 56 234	94.42	96.88	95.93	95.16
5 16 234	89.05	94.25	91.33	90.08
4 26 135	88.57	94.00	91.39	89.59
2 36 145	91.87	95.27	93.01	92.50
3 16 245	92.29	95.37	93.89	92.89
5 36 124	89.27	94.17	92.43	90.37
15 26 34	88.53	94.15	90.83	89.70
1 35 246	89.95	94.98	91.80	90.82
1 45 236	91.92	95.28	93.37	92.14
1 26 345	92.01	95.43	93.72	92.67
1 36 245	92.27	95.48	93.89	92.71
1 46 235	90.15	94.62	91.22	90.76
6 25 134	88.97	94.24	90.85	89.81
1235 4 6	89.18	94.06	90.75	90.04
1234 5 6	89.62	94.42	91.67	90.62
6 12 345	92.52	95.54	93.69	92.81
15 24 36	88.77	94.30	90.42	89.83
6 35 124	89.51	94.12	91.78	90.50
6 23 145	92.09	95.29	93.74	92.31
6 24 135	88.99	94.09	91.20	90.14
16 25 34	88.79	93.94	90.70	89.89
16 23 45	91.89	95.24	93.93	92.25
6 15 234	89.14	94.20	90.64	90.15
4 16 235	88.87	93.87	91.09	89.73
14 25 36	88.55	93.96	91.05	89.49
6 13 245	92.30	95.58	93.87	92.92
12 36 45	92.07	95.39	93.87	92.47
6 45 123	92.30	95.46	94.04	92.78
2345 1 6	92.75	95.66	93.78	93.48
16 24 35	88.88	94.18	89.76	90.16
1 25 346	90.09	94.94	91.74	90.70
6 14 235	88.83	93.86	89.99	89.86
4 36 125	88.97	93.99	90.53	89.74
6 34 125	89.28	94.28	90.75	90.10
1245 3 6	92.85	95.54	93.63	93.28
1345 2 6	92.20	95.47	93.09	92.94
2 16 345	91.96	95.42	93.57	92.55
2346 1 5	90.32	95.27	91.48	90.92

Tabela 4. Resultado da aplicação do Algoritmo 2

Referências

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Potts, S., Biesmeijer, J., Kremen, C., Neumann, P., Schweiger, O., and Kunin, W. (2010). Global pollinator declines: Trends, impacts and drivers. *Trends in ecology & evolution*, 25:345–53.

Potts, S. G. et al. (2016). Safeguarding pollinators and their values to human well-being. *Nature*, 540:220–229.