IoT for Development: Building a Classification Algorithm to Help Beekeepers Detect Honeybee Health Problems Early

Antonio Rafael Braga, Edgar E. Hassler, Danielo G. Gomes, Breno M. Freitas, Joseph A. Cazier

Outline

- Introduction
- Related works
- Materials and Methods
  - Dataset
  - Preprocessing
    - Exploratory Data Analysis (EDA)
    - Scatter plot
    - Basic Statistics
    - Detecting and removing outliers
    - Dada resizing (standardization)
- Machine Learning Strategies
  - Classification Algorithms
  - Evaluation Metrics
- Experiment Setup
- Results
- Conclusion

Introdução

- It is estimated that bees are responsible for a $18 billion in pollination service;
- Despite the growing number of colonies in the US, beekeepers lose an average of 30% of their colonies each winter;
- In Brazil, in recent years, Northeastern semi-arid beekeepers have suffered severe losses due to *absconding*.

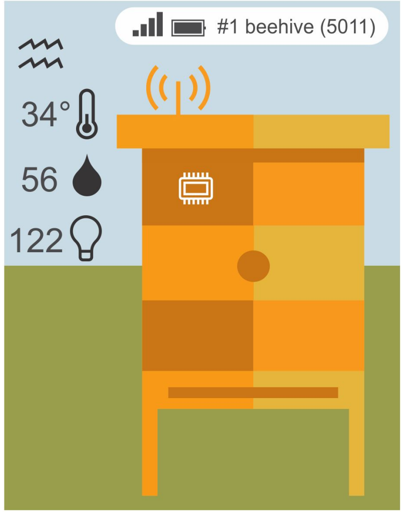
Fonte: http://g1.globo.com/natureza/noticia/2013/09/ciencia-busca-causa-de-colapso-das-colonias-de-abelha-no-mundo.html

Introdução

- To try to identify problems in the colonies in advance, the beekeeper usually performs a **inspection**. But:
  - stressful;
  - cause microclimate imbalance;
  - disturb the work of bees;
  - disrupts the organization of the colony;
  - can kill workers or even the queen crushed by removing and inserting the frames.

Introdução

- Sensors in the beehives can reduce the periodicity of physical management;
- Using the Internet of Things (IoT) it is possible monitorate:
    - temperature, humidity, carbon dioxide (CO2) and oxygen concentration, hive mass, image, sound, and so on.



- **Questão de pesquisa**
    - “Is it possible to identify health status of *Apis mellifera* colonies through sensors data (internal and external) and inspection data?”

Related Works

	Internal data	Internal and external data	Internal, external and inspection data
Kridi et al. 2014	✓	-	-
Kviesis and Zacepins 2016	✓	-	-
Carvalho et al. 2018	✓	-	-
Robles-Guerrero et al. 2019	✓	-	-
Maciel et al. 2018	✓	✓	-
Este trabalho	✓	✓	✓

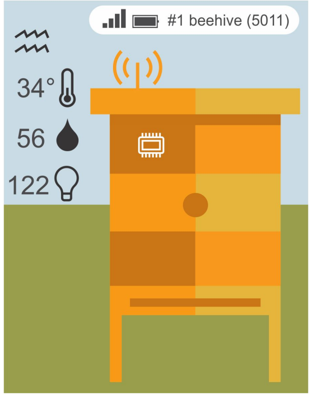
Table 1. Related works summary

Materiais e Métodos

- Dataset
  - Bayer Bee Care Center;
  - Located in Durham, Carolina do Norte, EUA;
  - Colony of *Apis mellifera;*
  - National Weather Service (NWS);
  - 6 months of monitoring;
  - Weekly inspection;

Sensors data

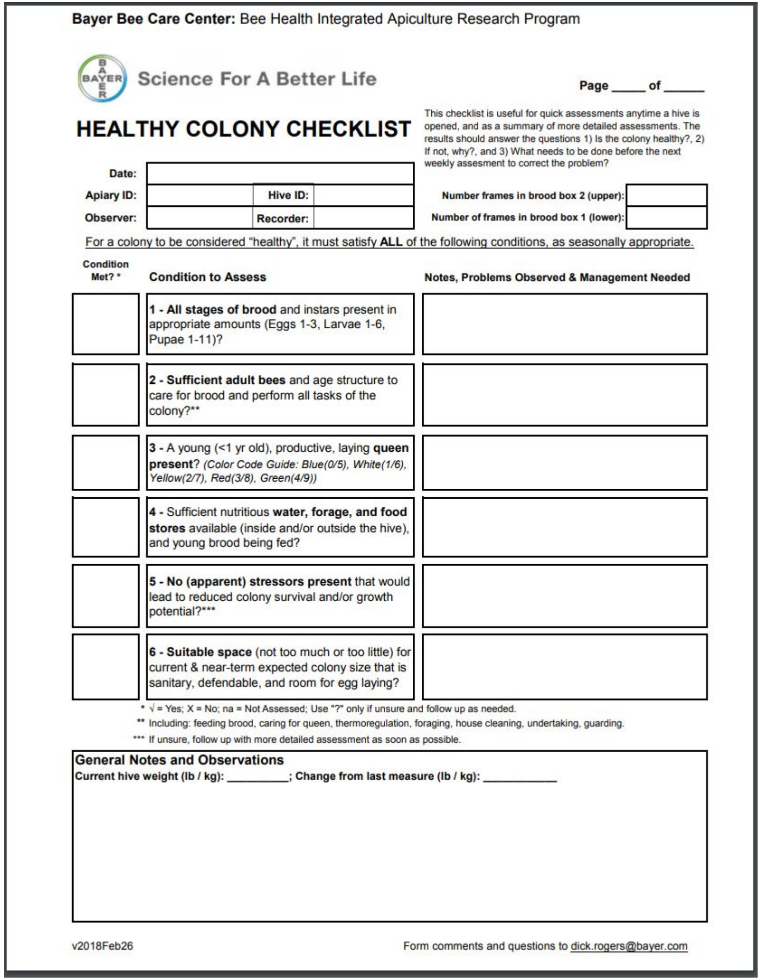
- Internal sensors
  - bees cluster temperature;
  - bees cluster humidity;
  - beehive temperature;
  - beehive humidity;
  - beehive weight.
- External sensores
  - external temperature;
  - dew point;
  - pressure;
  - wind speed;
  - precipitation per hour.



https://www.climaeambiente.com.br/prod,idproduto,3730043,estacao-meteorologica-estacao-meteorologica-vantage-pro2-davis--300-metros----rad--solar---k6162nouv

Healthy Colony Checklist

- 1 - All stages of brood;
- 2 - Sufficient adult bees;
- 3 - A young (<1 year old), productive, laying queen;
- 4 - Sufficient nutritious water, forage, and food stores;
- 5 - Presence of stressors;
- 6 - Suitable space.



Bayer Bee Care Center: Bee Health Integrated Apiculture Research Program

BAYER Science For A Better Life

Page _____ of _____

HEALTHY COLONY CHECKLIST

This checklist is useful for quick assessments anytime a hive is opened, and as a summary of more detailed assessments. The results should answer the questions 1) Is the colony healthy?, 2) If not, why?, and 3) What needs to be done before the next weekly assesment to correct the problem?

Date:

Apiary ID: **Hive ID:** **Number frames in brood box 2 (upper):**

Observer: **Recorder:** **Number of frames in brood box 1 (lower):**

For a colony to be considered "healthy", it must satisfy **ALL** of the following conditions, as seasonally appropriate.

Condition Met? *	Condition to Assess	Notes, Problems Observed & Management Needed
	1 - All stages of brood and instars present in appropriate amounts (Eggs 1-3, Larvae 1-6, Pupae 1-11)?	
	2 - Sufficient adult bees and age structure to care for brood and perform all tasks of the colony?**	
	3 - A young (<1 yr old), productive, laying **queen present**? *(Color Code Guide: Blue(0/5), White(1/6), Yellow(2/7), Red(3/8), Green(4/9))*	
	4 - Sufficient nutritious **water, forage, and food stores** available (inside and/or outside the hive), and young brood being fed?	
	5 - No (apparent) stressors present that would lead to reduced colony survival and/or growth potential?***	
	6 - Suitable space (not too much or too little) for current & near-term expected colony size that is sanitary, defendable, and room for egg laying?	

* √ = Yes; X = No; na = Not Assessed; Use "?" only if unsure and follow up as needed.
** Including: feeding brood, caring for queen, thermoregulation, foraging, house cleaning, undertaking, guarding.
*** If unsure, follow up with more detailed assessment as soon as possible.

General Notes and Observations
Current hive weight (lb / kg): __________; **Change from last measure (lb / kg):** ____________

v2018Feb26

Form comments and questions to dick.rogers@bayer.com

Health Status

- To calculate the colony health status, for each HCC feature was given binary value.
    - If **all** items **are ok** (receive value 1) the colony is classified as healthy, **class C3**;
    - If **1** or **2** items **are not ok** (receive value 0), the colony is classified às attention, **class C2**;
    - If **3** or **4** items **are not ok** (receive value 0), the colony is classified as not healthy, **class C1**;
    - And, if more than **5** items **are not ok** (receive value 0), the colony is classified as sick, **class C0**.

Preprocessing

- Exploratory Data Analysis (EDA)
  - mean, standard deviation, interquartile analysis and obliquity;
- Scatter plot

Preprocessing

- Basic Statistics
- Detecting and removing outliers (interquartile)
  - $[Q1 - 1.5 \times (Q3 - Q1),$
  - $Q3 + 1.5 \times (Q3 - Q1)]$
- Dada resizing (standardization)

$$X_{new} = \frac{x - \mu}{\sigma}$$

	Temp-Btm(F)	Temp-Brood(F)	Temp-Hive(F)	BRH(%)	HRH(%)	Weight(lbs)
count	180.000000	180.000000	180.000000	180.000000	180.000000	180.000000
mean	76.143333	87.266667	84.062778	58.772222	56.526111	47.710556
std	6.921548	7.086489	9.161134	6.393345	7.795260	7.219639
min	54.800000	71.900000	71.300000	27.000000	25.100000	27.900000
25%	72.650000	84.525000	72.700000	55.675000	53.400000	43.575000
50%	77.300000	89.200000	88.050000	60.100000	57.850000	46.100000
75%	81.025000	92.700000	92.700000	62.900000	61.850000	51.725000
max	88.100000	95.400000	95.100000	70.600000	69.800000	64.900000

Machine Learning Strategies

- Classification Algorithms
  - Naive Bayes (NB): is based on the Bayes Theorem to generate predictions for each observation by classifying a sample into a group that is most likely to have it;
  - k-Nearest Neighbors (kNN): The classification is then performed according to the highest number of k's closest neighbors belonging to the same class;
  - Random Forest (RF): is based on a group (or forest) of decision trees used to generate the classifications;
  - Neural Networks (NN): is based on a metaphor of the behavior of the brain.

Machine Learning Strategies

- Evaluation Metrics
  - Classification Accuracy (CA): is the accuracy of the classification model.

$$CA = \frac{vp + vn}{tp + tn + fp + fn}$$

  - Precision: expressed the proportion of correctly classified samples, considering the set of all samples classified (correctly and incorrectly).

$$precision = \frac{tp}{tp + fp}$$

  - Recall: or true positive rate explains how effectively the classifier identifies positive predictions.

$$recall = \frac{tp}{tp + fn}$$

  - F1-score: is a way of balancing Precision and Sensitivity.

$$F1\text{-}score = 2 * \frac{precision * recall}{precision * recall}.$$

tp = true positive, *tn* = true negative, *fp* = false positive, and *fn* = false negative

Experiment Setup

- Data Labeling
  - The classes were **obtained** through the inspection worksheet by overlapping process between sensed and inspection data.
  - The classes were **assigned** in the sensors samples subsequent to an inspection in such a way that it remained the same between one inspection and another until the state of the colony changed.
- Data Splitting
  - 60% to train, 20% to validation, and 20% to test with 10-fold cross-validation.
- Hyper-parameters
  - Naive Bayes: performed 50 times;
  - k-NN: k = 1;
  - Randon Forest: 76 trees;
  - Neural Network: architecture 1:5:2:1, i. e., 2 hidden layers (with 5 and 2 neurons).

Results

- NB and kNN Confusion matrices

	C0	C1	C2	C3
C0	**2**	0	0	0
C1	0	**5**	1	0
C2	0	2	**80**	15
C3	0	0	34	**41**

Table 4. NB Confusion Matrix

	C0	C1	C2	C3
C0	**2**	0	0	0
C1	0	**4**	1	1
C2	0	0	**78**	19
C3	0	0	19	**56**

Table 5. kNN Confusion Matrix

Results

- RF and NN Confusion matrices

	C0	C1	C2	C3
C0	**2**	0	0	0
C1	0	**4**	1	1
C2	0	0	**75**	22
C3	0	0	18	**55**

Table 6. RF Confusion Matrix

	C0	C1	C2	C3
C0	**0**	0	0	2
C1	0	**0**	0	6
C2	0	0	**78**	19
C3	0	0	14	**61**

Table 7. NN Confusion Matrix

Results

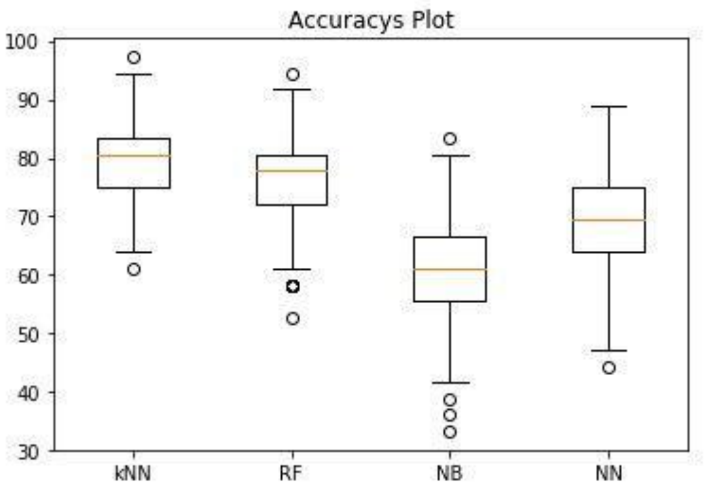
- Metrics

	CA	Precision	Recall	F1-score
NB	60.11%	0.59	0.72	0.65
kNN	**80.04%**	**0.80**	**0.80**	**0.79**
RF	78.20%	0.78	0.78	0.78
NN	77.22%	0.77	0.75	0.76

Tabela 8. Métricas de avaliação da acurácia dos algoritmos de classificação.

Results

- Accuracies Plot



Boxplot with the success rate of the classification algorithms.

Conclusion

- We present a machine-learning approach to predicting bee colony health status *Apis mellifera*;
- Were used data from internal and external sensors from a hive that was monitored for a period of 6 months;
- Was generated 4 classifications models from Naive Bayes, k-Nearest Neighbors, Random Forest e Neural Networks;
- We get hit rate that reached 80% with kNN algorithm;
- Thus, the answer for the research question is YES. So, from sensors and inspection data it is possible determine the health state of a bee colony *Apis mellifera* with high precision;

Questions?