Artificial Bee Colony Algorithm for Feature Selection in Fraud Detection Process

Gabriel Covello Furlanetto, Vitória Zanon Gomes, Fabricio Aparecido Breve

unesp

Fabricio Breve
São Paulo State University - UNESP
fabricio.breve@unesp.br

CAPES
FAPESP

Agenda

- Introduction
- Related works
- Artificial Bee Colony Algorithm
- The proposed work
- Results and discussion
- Conclusion

Introduction

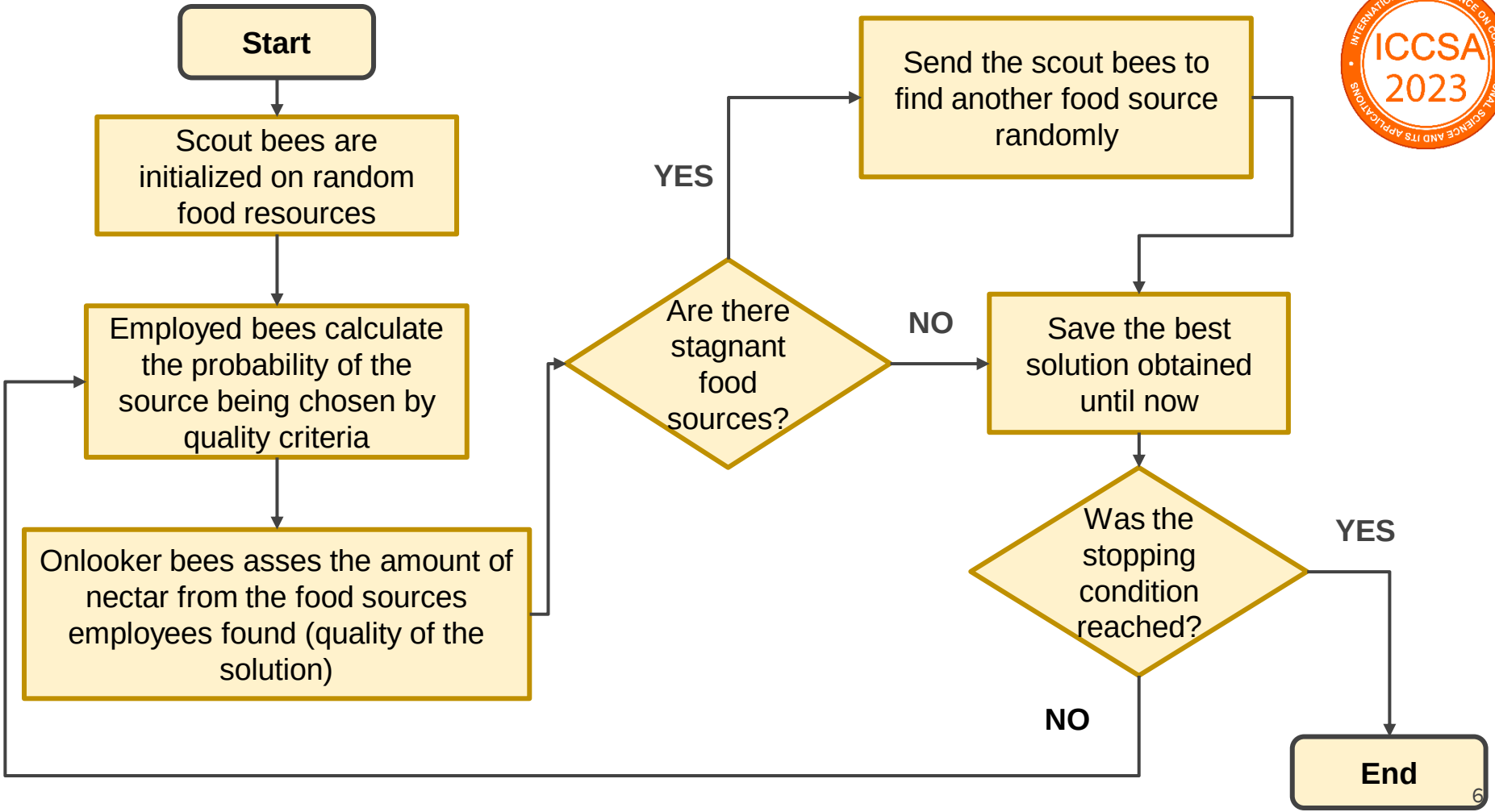
- Fraud detection became indispensable for companies
  - Increase of banking operations via smartphones and the internet.
  - Frauds grew during the COVID-19 pandemic.
- Machine learning techniques are widely used for this purpose
  - Some of those techniques use supervised learning and hybrid approaches.
- The datasets that collect financial transactions have to deal with a huge amount of attributes
  - They can be often irrelevant and/or redundant, making the search for patterns consume time and computational resources without getting relevant results.
  - To avoid that, Feature Selection techniques can be applied aiming to keep only the relevant and highly related ones.

Introduction

- In this work we use the Artificial Bee Colony (ABC) algorithm as a feature selection method and as an parameter optimizer to feed other feature selection algorithms such as K-best and RFE.
  - ABC is known by its simplicity and results quality.
  - It needs just a few parameters to work, unlike other similar algorithms.
  - It can be easily adapted for algorithm hybridization.

Artificial Bee Colony Algorithm

- Created by Karaboga (2005) initially for numerical optimization.
- Simulates the foraging process of honey bees.
  - The worker bees are divided into three groups.
    - Scout bees: randomly look for food sources.
    - Employed bees: responsible for one food source in specific, evaluating it and bringing a sample of the food to the hive.
    - Onlooker bees: summarize and evaluate all the information brought by the employees and decide which source should be exploited.
- Food sources represent possible solutions to the problem.
  - They are evaluated through an objective function.
- The algorithm works iteratively until a stopping condition is reached, and then the best solution found is returned.


Start
Scout bees are initialized on random food resources
Employed bees calculate the probability of the source being chosen by quality criteria
Onlooker bees asses the amount of nectar from the food sources employees found (quality of the solution)
Are there stagnant food sources?
YES
Send the scout bees to find another food source randomly
NO
Save the best solution obtained until now
Was the stopping condition reached?
YES
NO
End

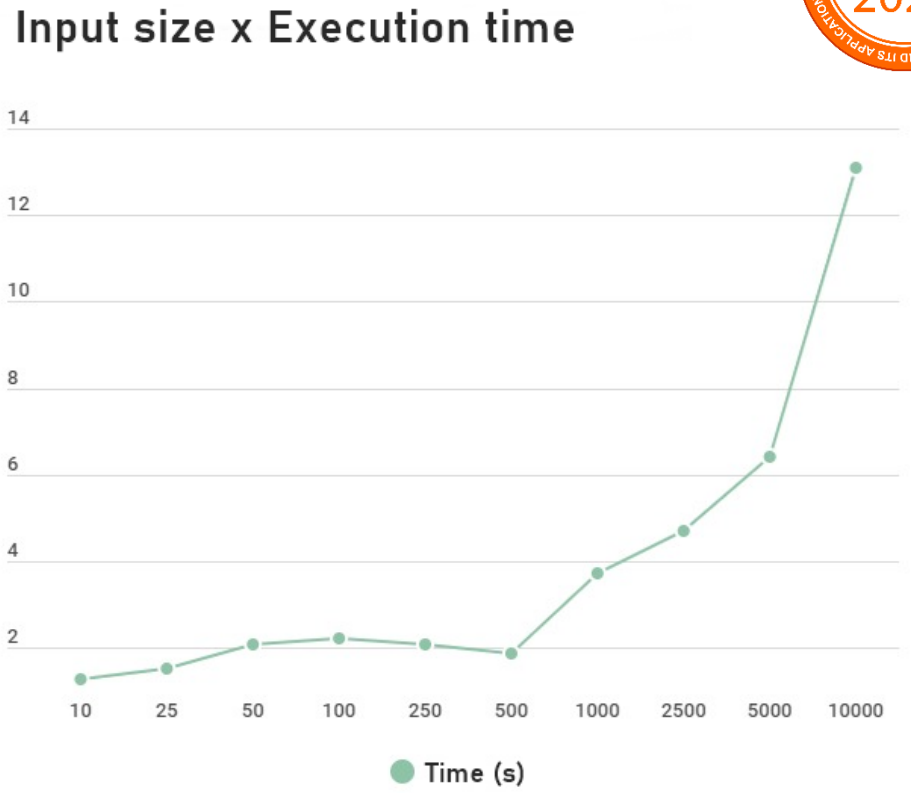
The proposed work

- First step: Look for a dataset to train, test and validate the model
  - Credit Card Transactions Fraud Detection Dataset[1].
- Second step: Implementing ABC using Python
  - Large number of libraries for machine learning and statistic.
  - Bee class.
    - Attributes: which food source is under the bee's responsibility, quality of the source and if it is stagnant or not.
- Objective function
  - Performs the train/test process through one of three different classifiers and evaluates the results through the F1-score accuracy measure.
    - Logistic Regression
    - Gradient Boosting
    - Random Forest
  - Therefore, ABC objective is to search for a set of attributes that reduce the final error and maximize the F1-score measure, leading to better classification results

1. https://www.kaggle.com/datasets/kartik2112/fraud-detection

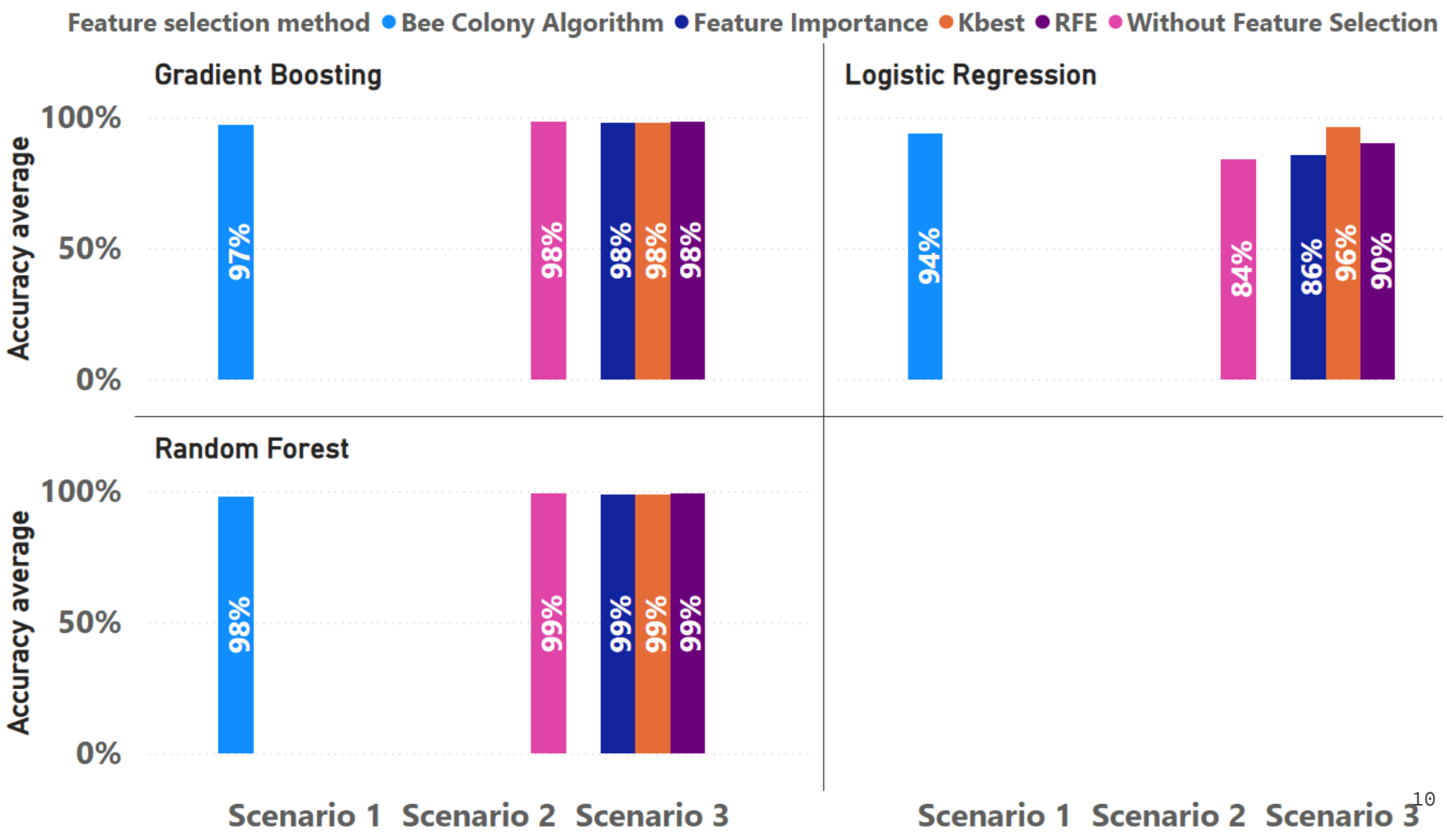
The proposed work

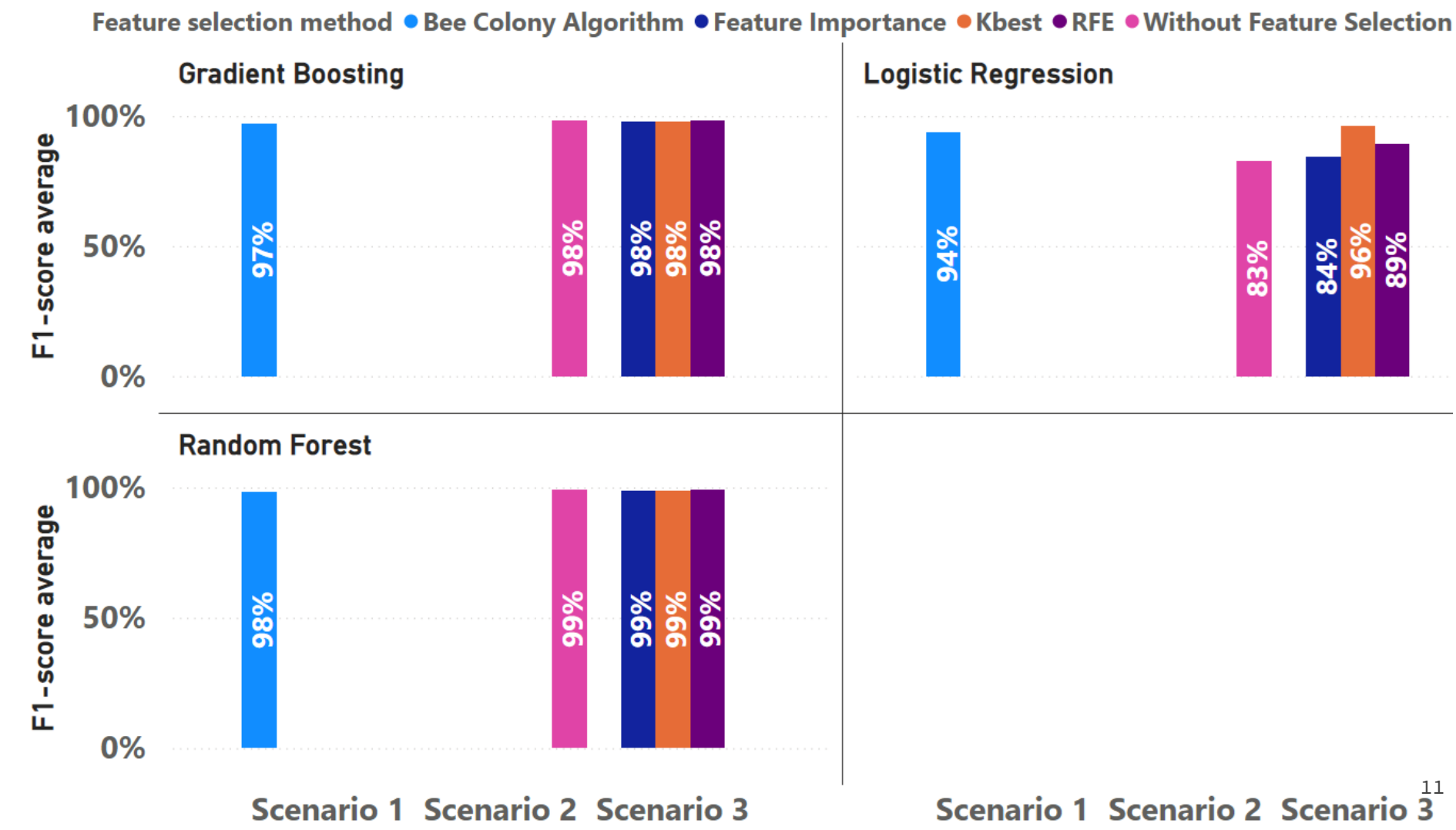
- ABC's computational complexity order: $O(n) = 2^n$
  - Feature selection, just like other optimization problems, is NP-complete.



Results and discussion

- Three scenarios proposed:
  1. ABC is used for feature selection and then Logistic Regression, Random Forest and Gradient Boosting algorithms are applied to classify the dataset using the set of attributes found by ABC
  2. Logistic Regression, Random Forest and Gradient Boosting algorithms are applied to the whole dataset, without a feature selection process.
  3. ABC is used to find the best amount of attributes and this value is used as input to K-best, RFE and Feature Importance algorithms which will perform the feature selection process and the same three classifiers used in scenarios 1 and 2 are applied to the new set of features.
- Each scenario was executed 100 times, and ABC parameters were:
  - Maximum number of iterations: 100
  - Number of bees: 10
  - Stagnation limit: 5


Feature selection method
Bee Colony Algorithm
Feature Importance
Kbest
RFE
Without Feature Selection
Gradient Boosting
Logistic Regression
Random Forest
Accuracy average
100%
50%
0%
97%
98%
98%
98%
98%
94%
84%
86%
96%
90%
98%
99%
99%
99%
99%
Scenario 1
Scenario 2
Scenario 3
Scenario 1
Scenario 2
Scenario 3


Feature selection method
Bee Colony Algorithm
Feature Importance
Kbest
RFE
Without Feature Selection
Gradient Boosting
Logistic Regression
Random Forest
F1-score average
100%
50%
0%
97%
98%
98%
98%
98%
94%
83%
84%
96%
89%
98%
99%
99%
99%
99%
Scenario 1
Scenario 2
Scenario 3

Results and discussion

- All tests were performed on a notebook with a 64-bit Windows 10 operating system, 16.0 GB RAM, Intel(R) Core(TM) i7-1165G7, 2.80GHz, of 11th generation.
- 2 different situations observed:
  - For the Gradient Boosting and Random Forest classifiers, the use of the ABC algorithm practically does not change the accuracy and the F1-Score of the results.
    - The biggest difference is presented when there is a direct application of the algorithm to select attributes (Scenario 1), in which the results end up getting worse by approximately 1%.
  - For the Logistic Regression classification model, there is an increase in performance, both for the accuracy and the F1-Score, for Scenarios 1 and 3 with respect to Scenario 2 (no attribute selection).

Conclusion

- This article presented a new feature selection technique based on the ABC algorithm.

- The results show reduction of database complexity, and achieve a higher accuracy of classification than that obtained when using the complete set of data.

- The results obtained corroborate the quality of ABC as a wrapper method for feature selection, validating the hypothesis that the algorithm could reduce the cost and increase the quality of the results obtained in a fraud detection process, so significant nowadays.

Conclusion

- The results are comparable to other state-of-art methods, a fact that encourages its use for such tasks.

- ABC can also be used to optimize the parameters from other feature selection methods, a secondary task that heavily influences the quality of the result.

Artificial Bee Colony Algorithm for Feature Selection in Fraud Detection Process

Gabriel Covello Furlanetto, Vitória Zanon Gomes, Fabricio Aparecido Breve

unesp

Fabricio Breve
São Paulo State University - UNESP
fabricio.breve@unesp.br

CAPES
FAPESP