Smart Staffing

Machine Learning for optimizing crew scheduling

TAIBAOUI Mohammed and[1] TALMAT Amin Mohamed[1]

Supervised by: REZKI Nafissa[1]

Abstract

The problem of crew scheduling in airlines has received considerable attention in recent years. This problem is often broken down into two steps due to its large size and increasing complexity, first crew pairing, then crew assignment. This work focuses on the crew pairing problem, also known as "CPP ", which is an essential part in crew planning and consists of creating sequences (pairings) of flights, where the company Air Algeria encounters difficulties which require a lot of time to solve this problem in a feasible manner. The objective is to apply machine learning models in solving this problem in order to optimize the use of crews. At first, we tried some machine learning methods, but the results obtained weren't as good, so we used the Machine Learning methods to reduce the complexity of the problem instead by splitting the data, then suggested an algorithm that gave much better and feasible results.

Keywords: Machine Learning, crew pairing problem, clustering, optimization

1.Introduction

Crew Scheduling is very important for airline companies, given that the costs related to crews ranks second after fuel [1]. Also mentioning that the industry is growing by size and volume, the scheduling becomes more and more complex with the flowing of time. It is then necessary to develop new algorithmic methods that meet the problem which doesn't cease to grow.

This work focuses on the crew pairing problem, which is basically creating rotations or duties that can be done by the same crew, and meets certain requirements and constraints related to them, mentioning that all flights have to be covered.

The objective of this work is to use Machine learning methods and techniques during the process of solving this problem to reduce the time taken by this phase of scheduling for the company.

In the course of this work, we tried to apply several ML approaches such as the ANN approach but the results were not of good quality due to the incompatibility of the problem data with the ANN intuition, so we used the clustering approach to reduce the problem complexity and suggested a method to generate results close to a real result. Where the process takes less time and generates fewer additional costs (off-base time).

---

[1] *Department of Logistics and Transport Engineering, National Superior School of Advanced Technologies (ENSTA) Dergana, Algeria.*

2.Basic concepts:

Before starting this work, it is required to provide some context about AI, Machine Learning and the work environment for better understanding.

2.1 Artificial Intelligence:

Artificial intelligence is a vast field in computer science that has been created to perform services and tasks requiring human intelligence. Its aim is to develop machines capable of adapting to their environment, and reasoning of making decisions, as it refers to machine learning, deep learning, artificial neural networks and reinforcement learning, among others.

2.2 Machine learning:

Machine Learning is a developing technology and a particular branch of artificial intelligence that enables computers to learn autonomously from historical data, i.e. instead of taking a computer to perform a given task, it is provided with data and algorithms that enable it to learn from this data to perform the task itself and make decisions for new data from their past experiences by recognizing patterns.

2.3. Air Algérie

Air Algérie is a public economic enterprise organized as a joint stock company (SPA) with a capital of 60,000,000,000 DZD. It provides transport for over 5 million passengers (2019) to 77 destinations (4 continents, 44 international and 33 domestic services) with a fleet of 56 aircraft that meet international safety standards [2].

In Air Algérie, the scheduling is mostly done manually, which is neither effective nor practical in terms of optimization of exploiting the crews without mentioning how much time it takes. The literature review proved that studies have been done about Machine Learning in the airline industry, and also proved that implementing it had significant improvements.

2.4. Airline resource planning:

The organization and planning of aircraft and crews within airlines is an extremely complex task, and is therefore generally divided into several stages, with the result of each stage serving as input for the next.

First of all, the schedule is drawn up to match the marketing department's expectations with the available fleets and network constraints, such as the slots available to the airline at the various airports.

Secondly, in the stage of assigning aircraft to flight segments, taking into account the days on which they are operational and those reserved for maintenance, since the revenue forecast for a flight stage depends on the type of aircraft used, there are airports where certain aircraft are unable to take off or land.

Then, once the aircraft are assigned, it begins the pairing operation, which consists of a series of flight legs (flight sequences) for an unassigned crew member, with take-off and landing at the same crew base. During these sequences, work is generally grouped into duty slots, and these pairings must comply with a large number of government laws and collective agreements that differ from one airline to another.

Finally, to complete resource planning by assigning crews, the last step is to assign pairs of individuals (cabin crew and flight crew, including pilots, co-pilots, stewards and stewardesses) in the aim is to ensure coverage of all flight sequences with equal working hours between the different crews, while satisfying work rules and regulations.

3.Evolution of approaches to crew planning:

From 1950 to 1990, the air transport industry used operational research techniques such as integer programming and linear relaxation programming, which consists of simplifying CPP into a linear program that is easier to solve. But with the growth in data size and air networks, the complexity of this problem is increasing, making these techniques unusable and pushing researchers to look for new approaches.

From 1990 onwards, the emergence of new techniques based on column generation such as Branch-and-Price, which combines column generation with a Branch-and-Bound technique for integer programming [3]. For example, in 1997 Barnhart et al [4] explored an approach to identifying near-optimal solutions to the crew pairing problem, by combining a dynamic column generation with a customized Branch-and-Bound procedure.

In the early 2000s, the emergence of heuristics and metaheuristics - approaches that involve finding good solutions in a short time - and the rise of machine learning (ML) techniques led to several works that marked an improvement in solving this problem. Among these works are Yaakoubi et al. (2020)) [5] use machine learning to create initial solutions and reorganize the clusters of the Crew pairing problem (CPP) based on column generation, in parallel Morabit et al. (2021) [6] employ a graphical neural network to choose the pairs to be preserved during the generation of which the aim is to decrease the time spent on the solver, Quesnel et al. (2022)) [7] predict the matches will be beneficial by filtering the matches during the generation of columns of a crew assignment. The table below presents a summary of work in the airline industry, detailing the techniques/algorithms used, their contributions and results.

Table 1: Review of related work

Réf	Technique /Algorithme	Contribution	Results
[5]	- Clustering - Column generation - Constraint aggregation - ML-based heuristics	This paper presents a new method for solving CPP (crew pairing problem) by combining an improved basic solver with dynamic control. -flight grouping based on machine learning -advanced operational research techniques to assemble optimal solutions - He started with a basic solver and developed it further, incorporating artificial intelligence methods.	The article states that the Commercial-GENCOL-DCA method outperforms the BASELINE base solver in terms of performance for large-scale CPPs, with an average reduction in solution costs of between 6.8% and 8.52%, this reduction being mainly attributable to a significant drop (from 69.79% to 78.11%) in the costs of overall regulatory and crew limitation constraints.
[8]	- Artificial neural networks (ANN) - Adaptive neuro fuzzy inference systems (ANFIS)	- Evaluation of the performance of several machine learning algorithms compared to the traditional multiple linear regression model in the problem of modeling to forecast air travel demand applied on the airline to Australia.	- The ANFIS and ANN models have a lower RMSE, while the GA model has a higher RMSE. - each company has its own approach that suits it.

	-Genetic Algorithms (GA) -Support Vector Machines (SVM) -Regression Trees -Les test ANOVA	- Identify the approach that offers the greatest statistical accuracy - Integrating machine learning into air travel demand management.	- The study also established that AI-based methods can be successfully implemented in the aviation industry to support business planning and management.
[9]	- Decision tree, - Artificial Neuron Network - Support Vector Machine (SVM), - Random Forest - Markov chain	- Review flight crew capabilities before takeoffs. - Thanks to this integrated approach, not only is flight data classification improved, but it also provides a framework for improving cabin crew selection and route planning systems. - Proactively forecast cabin crew performance to reduce flight safety hazards and increase overall operational efficiency.	- The use of several machine learning models offers a global method for evaluating cockpit crew performance, enabling a finer understanding of performance factors. - Improve flight safety through detailed prediction of flight crew performance.
[10]	- multinomial logistic regression - decision trees - Clustering algorithm	- prediction of airlines' route choices between two airports based on the state of the two-airport cluster, according to the characteristics of each route, such as flight efficiency, air navigation charges and traffic forecasts. - the proposed approach offers significant potential for improving pre-tactical traffic forecasts, in order to optimize traffic flows to match available capacity	- the performance of decision tree regression is inferior to that of multinomial regression in certain situations. - The absence of certain crucial explanatory variables, such as the availability of each route choice, can influence the results.

4.Proposed Solution:

After mentioning how the size of the problem is an issue, we used a clustering method which is hierarchical clustering to reduce the complexity of the problem and proposed an algorithm to solve it which we are going to discuss its result after.

4.1 Clustering:

Using a clustering method to generate clusters or sets in such a way that each flight that belongs to a certain set is closer to the rest of the same set than flights in other sets, the hierarchical clustering approach will be adequate to do this since the intuition of this approach is to group similar objects in the same group based on variables presented by calculating the distances between these variables one by one and determining the optimal grouping, But first we need to use the dendrogram method to determine the optimal number of clusters, and here there are several linking methods and the most suitable for the problem is Ward's method because it minimizes the increase in total variance within the cluster after merging two clusters. And so it's suitable when you want to create balanced clusters with minimal variance. And for the parameter used we have:

n clusters: this is the number of desired clusters, determined using dendrograms.

Affinity: this is the metric used to calculate the linkage between clusters; here, the Euclidean metric will be the most appropriate, since we want to cluster based on straight-line distance.

Linkage: this is the linking method and will be the same as above for the same reasons, 'ward'.

And after model creation, we'll present our data for model training and receive the clustering result.

4.2 Implementation:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
dataset= pd.read_csv("/content/Flights.csv")
dataset['STD'] = pd.to_datetime(dataset['STD'])
dataset['STA'] = pd.to_datetime(dataset['STA'])
dataset['heure_depart_hours'] = dataset['STD'].dt.hour + dataset['STD'].dt.minute / 60
dataset['heure_arrive_hours'] = dataset['STA'].dt.hour + dataset['STA'].dt.minute / 60
dataset['STD'] = dataset['STD'].dt.time
dataset['STA'] = dataset['STA'].dt.time
dataset['depart_hours'] = (dataset['daydep']-1) * 24 + dataset['heure_depart_hours']
dataset['arrive_hours'] = (dataset['dayarr']-1) * 24 + dataset['heure_arrive_hours']
x = dataset[['depart_hours', 'arrive_hours']]
import scipy.cluster.hierarchy as sch
dendogram = sch.dendrogram(sch.linkage(x, method= 'ward'))
plt.title('Dendrogram')
plt.xlabel('flights')
plt.ylabel('ED')
plt.savefig('my_plot.png')
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters=2 , affinity = 'euclidean', linkage = 'ward')
y_hc = hc.fit_predict(x)
```

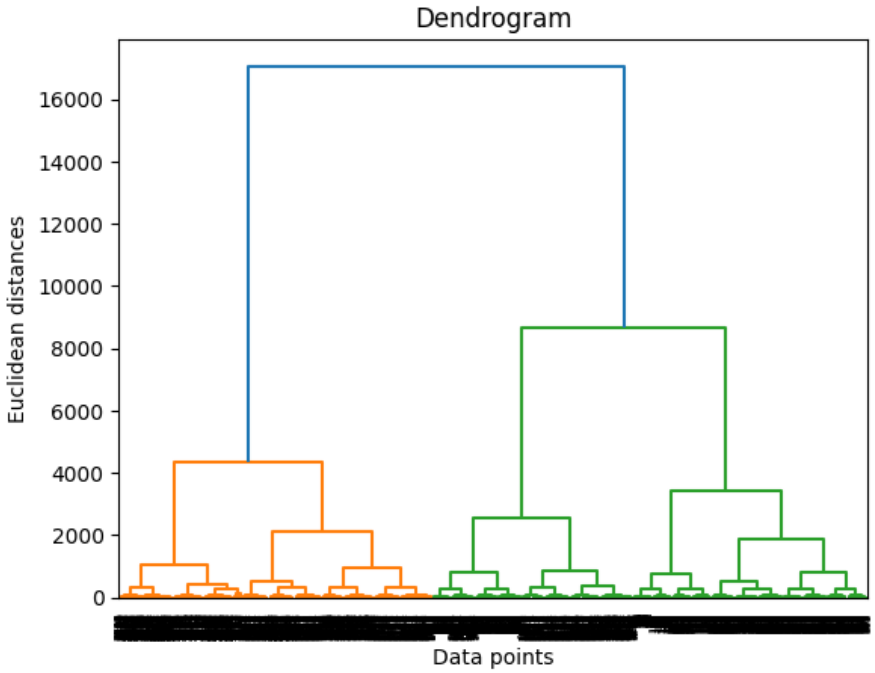

Figure 1: The Dendrogram

X-axis (DataPoints): The x-axis represents the individual flights that are grouped together. Each point on the x-axis represents one flight.

Y-axis (Euclidean distances): The y-axis, which stands for 'Euclidean distance', indicates the distance between data points in the feature space considered by the algorithm.

Horizontal links: Horizontal links represent the clusters formed at each stage of the algorithm. The height of the links indicates the distance between clusters; higher links represent greater distances.

We can see that the recommended number of clusters here is 2 if we cut the longest vertical line between two horizontal lines, and this is also indicated by the number of colors. So our data, which consists of 2391 lines (flights), will be divided into two sets of sizes 1497 and 894 flights.

4.3. The proposed algorithm:

The flight database is imported in semicolon csv format, an 'ID' column is added which is virtually an identification for the flights to facilitate handling and display of the results. Next, we'll classify the flights into 3 classes according to departure time;

B2nb: Flights departing from a base airport to a non-base airport.

Nb2b: Flights departing from a non-base airport to a base airport.

Nb2nb: Flights departing from a non-base airport to a non-base airport.

The sequence of flights will start from the base (b2nb) then the code will iterate through the flights from (nb2nb) then (nb2b) looking for the next possible flight that satisfies the constraints that are expressed as conditions (if clauses), if the flight respects the constraints their ID will be added to a list and the code will continue until it reaches 4 steps or after iterating through

all the flights. We've also added a Boolean column 'not_covered' initialized as true to track flight coverage, such that the value is set to False when the flight is covered. Sequences will be imported into a new data frame where the sequence departure airport, sequence arrival airport, departure and arrival time and sequence period are displayed.

After this step the code will check if there are any feasible sequences that start from base and return to base (rotation) that will be marked as ready and classify the rest into 3 sets, sequence starts from base and ends off-base, sequence starts off-base and ends on-base and sequence starts off-base and ends off-base, and flights not covered in the previous step are imported into the data frame and considered as a sequence that consists of a single flight.

Continue to the next step. Where we'll introduce 'layovers', which means that the Crew will spend the night away from its mother base. This will allow us to create more rotations, as one constraint will be changed for another, giving more freedom and possibility to the code, but it will create an additional cost, which is the hours away from the base.

And finally, we'll export the results to Excel after processing and organizing them in a data frame. And for flights and sequences not covered, we'll export them to Excel as well, in a separate file for manual processing.

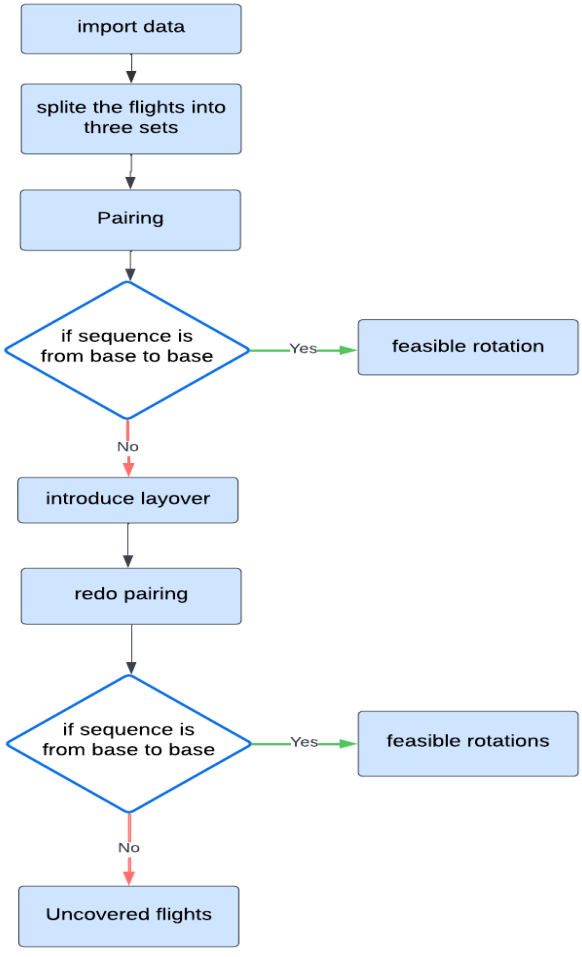

Figure 2: Descriptive diagram of the algorithm

5. Final result display and discussion

daydep	dayarr	Flight	DEP	ARR	STD	STA
1	1	1208	ALG	CDG	5:35:00 AM	8:00:00 AM
1	1	1209	CDG	ALG	9:00:00 AM	11:15:00 AM
/	/	/	/	/	/	/
1	1	6212	ALG	QSF	6:00:00 AM	6:30:00 AM
1	1	1108	QSF	ORY	7:30:00 AM	9:50:00 AM
1	1	1109	ORY	QSF	10:50:00 AM	1:10:00 PM
1	1	6213	QSF	ALG	2:10:00 PM	2:40:00 PM
/	/	/	/	/	/	/
1	1	6190	ALG	CZL	6:20:00 AM	7:10:00 AM
1	1	6191	CZL	ALG	8:10:00 AM	9:00:00 AM
/	/	/	/	/	/	/
1	1	6040	ALG	BSK	7:20:00 AM	8:10:00 AM
1	1	1124	BSK	ORY	9:10:00 AM	11:45:00 AM
1	1	1113	ORY	BJA	12:35:00 PM	2:50:00 PM
1	1	6051	BJA	ALG	3:50:00 PM	4:20:00 PM
/	/	/	/	/	/	/
1	1	6050	ALG	BJA	7:30:00 AM	8:00:00 AM
1	1	1112	BJA	ORY	9:00:00 AM	11:20:00 AM
1	1	1125	ORY	BSK	12:45:00 PM	3:15:00 PM
1	1	6041	BSK	ALG	4:15:00 PM	5:05:00 PM

Figure 3 : Screenshot of the results

The code generated 750 Crew Routes (669 without layovers) with 511,41667h off-base and 400 flights unaffected. After we finished, we compared the results of our work with the work done by the company:

Table 2: Table of results comparison

	Company	Algorithme
Number of rotations	747	750
Off-base hours	+5000	511.4167h
Number of flights not covered	0	400
Time taken	5 to 10 days	Less than 5min

Reduction in off-base hours: as a first observation, we note that our optimization algorithm generates rotations with significantly lower off-base hours than rotations created manually by the company, which implies the importance of using optimization algorithms in this problem, and that it will translate into lower costs linked to meal and hotel allowances for crews, as well as improved comfort for pilots and flight attendants.

Need for human intervention for uncovered flights: As a second remark concerning the algorithm, we have seen that 400 flights out of 2391 flights are not covered or not assigned to rotations, which requires human intervention because these flights are constrained by problems such as the flight ALG to ABJ, which will take 5 days to return to Algeria, at company level. In the practical case, the company uses solutions such as irregular flights or transporting crews as passengers to ensure continuity of service.

Saving time in execution: as a final remark on the time taken of the pairing operation, the algorithm considerably reduces the execution time compared to the manual method, allowing us to say that we have achieved the goal of reducing the processing time for our problem. In conclusion, the optimization algorithm offers several advantages over the manual method of creating air crew rotations in terms of reducing off-base hours and associated costs, improving crew comfort and saving time in executing the pairing operation. However, the algorithm fails to cover all flights, requiring human intervention for unassigned flights.

6.Conclusion

In the course of this work, we used the clustering approach to reduce problem complexity and suggested a method for generating results close to a real one. Where the process takes less time and generates fewer extra costs (off-base hours).

As for the limitations of our work, our method does not take into account the constraint linked to take-off and landing times, so in our work we have defined ALG airport as the only base because of the large number of flights that take off and land at this airport, our method only deals with one type of aircraft, and the results are not final results and they can be improved.

Consequently, execution time can be considerably improved by varying or modifying the algorithm responsible for constructing crew pairings. This algorithm can be used in conjunction with other algorithms in future work to enable improved algorithmic efficiency on monthly pairing problems.

Références

[1] D. Aggarwal, D. K. Saxena, M. Emmerich et S. Paulose, «On Large-Scale Airline Crew Pairing Generation,» *IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 593-600, 2018.

[2] Air Algérie, *Document interne*, 2023.

[3] Y. K. S. D. K. S. Divyam Aggarwal, «On Learning Combinatorial Patterns to Assist Large-Scale Airline Crew Pairing Optimization,» 2020.

[4] E. L. J. G. L. N. M. W. P. S. P. H. V. Cynthia Barnhart, «Branch-and-Price: Column Generation for Solving Huge Integer Programs,» *Operations Research*, vol. 46, n° %13, pp. 293-432, 1998.

[5] F. S. L.-J. Yassine Yaakoubi, «Machine learning in airline crew pairing to construct initial clusters for dynamic constraint aggregation,» *EURO Journal on Transportation and Logistics*, vol. 9, n° %1100020, 2020.

[6] G. D. ,. A. L. Mouad Morabit, «Machine-Learning–Based Column Selection for Column Generation,» *Transportation Science*, vol. 55, n° %14, pp. 815-967, 2021.

[7] A. W. G. D. F. S. Frédéric Quesnel, «Deep-learning-based partial pricing in a branch-and-price algorithm for personalized crew rostering,» *Computers & Operations Research*, vol. 138, 2022.

[8] G. B. P. S. S. R. Graham Wild, "Machine Learning for Air Transport Planning and Management," AIAA, 2022.

[9] F. J. M. T. Naimeh Borjalilu, "COCKPIT CREW SAFETY PERFORMANCE PREDICTION BASED ON THE INTEGRATED MACHINE LEARNING MULTI-CLASS CLASSIFICATION MODELS AND MARKOV CHAIN," Aviation, vol. 27, no. 3, p. 152–161, 2023.

[10] O. G.-C. R. H. Rodrigo Marcos, "A Machine Learning Approach to Air Traffic Route Choice Modelling," 2018.