Hybrid Estimation of Distribution Algorithm for solving a Resource Level Allocation Problem in a Legal Business

Mayowa Ayodele
The University of Manchester
Manchester, United Kingdom
mayowa.ayodele@manchester.ac.uk

K.Nadia Papamichail
The University of Manchester
Manchester, United Kingdom
n.papamichail@manchester.ac.uk

Geraldine Gallagher
DWF Law LLP
Manchester, United Kingdom
geraldine.gallagher@dwf.law

Darren Buckley
DWF Law LLP
Manchester, United Kingdom
darren.buckley@dwf.law

ABSTRACT

Resource level allocation entails assigning execution times to a set of resource levels required to complete a task. This is an important problem, particularly in the services sector. We consider a real-world variant of this problem originating from a legal business. The objective considered is the maximisation of damages savings.

We apply an Estimation of Distribution Algorithm (EDA) to this problem and use a machine learning model, Random Forest, as a fitness approximation method. The hybrid EDA presents promising results.

CCS CONCEPTS

•**Computing methodologies → Discrete space search;**

KEYWORDS

Estimation of Distribution Algorithm, Genetic Algorithm, Machine Learning, Fitness Approximation, Insurance, Legal Business, Random Forest, Resource Level Allocation

ACM Reference format:
Mayowa Ayodele, K.Nadia Papamichail, Geraldine Gallagher, and Darren Buckley. 2019. Hybrid Estimation of Distribution Algorithm for solving a Resource Level Allocation Problem in a Legal Business. In *Proceedings of the Genetic and Evolutionary Computation Conference 2019, Prague, Czech Republic, July 13–17, 2019 (GECCO '19),* 2 pages.
DOI: https://doi.org/10.1145/3319619.3326782

1 BACKGROUND

For many years, there has been research interest in combining Machine Learning (ML) with Evolutionary Algorithms (EAs). Some previous works focused on using EAs to improve ML models while others focused on using ML to improve EAs. ML models have particularly been used to enhance EAs in *Population Initilisation*, *Selection and Fitness Computation*, *Population Reproduction and Variation*, *Algorithm Adaptation* and *Local Search* [4].

GECCO '19, Prague, Czech Republic
 978-1-4503-6748-6/19/07…$15.00
DOI: https://doi.org/10.1145/3319619.3326782

Selection and Fitness Computation is the focus of this study. This category includes modelling objective fitness functions and reducing the number of fitness evaluations. In real-world problems, objective functions are often either expensive or impossible to compute analytically [4]. It is often necessary to create an approximation of the fitness. Some of the approximation models that exist in the literature are polynomial models, kriging models, Artificial Neural Network and Support Vector Machine [3]. In this paper, we use an ML model, namely Random Forest, to approximate the fitness function.

Estimation of Distribution Algorithms (EDAs) have presented promising results in many categories of optimisation problems such as the Permutation Flowshop Scheduling Problem [1] and the Quadratic Assignment Problem [2]. We will therefore explore the fitness approximation method within the context of an EDA.

The rest of this paper is structured as follows. A brief description of the Resource Level Allocation Problem (RLAP) is presented in 2. Section 3 presents the solution approach. Experiments and results are presented in 4 and 5. Conclusions are presented in 6

2 RESOURCE LEVEL ALLOCATION PROBLEM

Davies Wallis Foyster (DWF) Law LLP is a global legal business. Core UK Insurance is a significant part of the firm's portfolio. DWF acts for its clients; insurance companies, brokers, and self-insured when claims are made against them by a third party. DWF acts as the defendant solicitor and assesses the validity (liability) and value (quantum) of the claim from the third party. One of the desired outcomes for DWF is to minimise damages paid out by insurer clients.

From historic data, we see that the damages paid depends on some of the feature such as the time allotted to resource levels. To determine the optimal resource level allocation, we apply an ML model to predict the likely damages savings.

In this paper, we propose the MORLAP which is defined as follows. Given a job type, there exists at least one resource level $i \in [1, n]$ that can perform the job. Each resource level i requires a time t_i to complete the job. The aim is to find a set of time allocation to resource levels $\{t_1 \cdots t_n\}$ such that the damage savings ds is maximised.

The resource levels considered in ascending order of expertise are *Support*, *Paralegal*, *Trainee*, *Solicitor 0-2 yrs*, *Solicitor 2-4 yrs*, *Senior Solicitor*, *Senior Associate*, *Director*, *Partner* and *Senior Partner*.

3 SOLUTION APPROACH

A solution to the MORAP is represented as a string of integers where each index represents a resource level and each integer value represents the units of time spent on a claim. In this paper, one unit of time is set as 6 minutes. This is because staff time is recorded as a multiple of 6 minutes. The maximum unit of time is set to 315 units (31.5hrs), this is the maximum time spent on a matter in the last 1 year. The solution length is set to 11 which is the number of resource levels used in this paper.

The hybrid EDA used in this study is described in Alg. 1.

Algorithm 1 Hybrid EDA for MORLAP

```
Build Predictive Model M_pred
Initialise ts, ps and gen
Initialise Probabilistic Model M_prob(0)
for g = 1 to gen do
    Generate population P of size ps by sampling M_prob(g − 1)
    Predict dp of individuals in P using M_pred
    Assign fitness to individuals in P
    Set b as the best solution in P
    Select best ts < ps solutions to form S
    Generate M_prob(g) using S
end for
return b
```

In 1 of Alg. 1, the Random Forest model is used to generate M_{pred}. Note that M_{pred} is only built once and at the start of the algorithm. The parameters of the hybrid EDA which are truncation size ts, population size ps and the number of generations gen are then initialised. Note that gen is calculated as the number of fitness evaluations (ne) divided by ps.

Note that the probabilistic model M_{prob} is a matrix where each value p_{ij} is the probability of assigning i units of time to resource level j. $M_{prob}(g)$ is used to denote M_{prob} generated at generation g. At each generation, a population of solutions P is generated by sampling $M_{prob}(g-1)$. The fitness is set to $\%ds \div noOfLevels$ where $\%ds = ds \div dc$, dc is used to denote damages claimed. With this approach, the number of non-zero resource level allocation time, $noOfLevels$ is used to penalise the fitness of the solution. This will bias the search towards solutions that use less number of resource levels. Once the fitness of all solutions have been estimated, a population of ts best solutions is then selected from P to generate S. The population of promising solutions S is then used to build the probabilistic model at generation g, $M_{prob}(g)$ and completely replaces $M_{prob}(g-1)$. This process is repeated for gen generations after which the best solution b is returned.

4 EXPERIMENTAL SETTINGS

The scikit-learn[1] implementation of Random Forest is used in this paper. The number of trees is set to 200 while the maximum depth is 6. Default settings are used for all other parameters. The performance measure used to determine the prediction accuracy is r^2 which is the square of the sample correlation coefficient between the observed outcomes and the observed predictor values. We use the well-known 10-fold cross validation approach.

The features used to build the Random Forest model are *Client*, *Damages Claimed*, *Type of Claim*, and *Resource time*. These are features with higher correlation with the damages paid. To demonstrate the value in the proposed approach and since the only parameters the business has control over is the set of resource level time allocation, we only search the space of solutions for resource level time allocation.

For the EDA, we respectively set the number of evaluations (ne), number of runs (nr), population size ps and truncation size (ts) to 2000, 20, 100 and 10. This was chosen based on preliminary experiments suggesting good performance.

5 RESULTS AND ANALYSIS

For the Random Forest, the average r^2 is 0.752. The top three results generated at the final generation of the 20 runs of the algorithm are presented in Table 1. We see that a higher amount of resource leads to a higher prediction of damage savings. The top three solutions respectively assigned 26.9, 25.1 and 17.6 resource time to paralegals. This also is realistic as the case study problem is a low-value work which is often assigned to paralegals.

Table 1: Top three solutions

Fitness	%ds	Solution									
56	56	0	269	0	0	0	0	0	0	0	0
50	50	0	251	0	0	0	0	0	0	0	0
43	43	0	176	0	0	0	0	0	0	0	0

6 SUMMARY

A new class of allocation problem, MORLAP is formulated in this paper. This is a real-world problem with high relevance in legal businesses. The Random Forest has been used to predict the fitness of a solution in an EDA. The overall approach is the main contribution of this paper. Other ML models or evolutionary algorithms can be combined in a similar manner in the future.

In this future, to make this solution more realistic, we will consider the multi-objective version of this problem where it is important to also minimise cost as well as maximise damage savings.

REFERENCES

[1] Mayowa Ayodele, John McCall, Olivier Regnier-Coudert, and Liam Bowie. 2017. A Random Key based Estimation of Distribution Algorithm for the Permutation Flowshop Scheduling Problem. In *Evolutionary Computation (CEC), 2017 IEEE Congress on.* IEEE, 2364–2371.

[2] Josu Ceberio, Ekhine Irurozki, Alexander Mendiburu, and Jose A Lozano. 2014. Extending distance-based ranking models in estimation of distribution algorithms. In *Evolutionary Computation (CEC), 2014 IEEE Congress on.* IEEE, 2459–2466.

[3] Yaochu Jin. 2005. A comprehensive survey of fitness approximation in evolutionary computation. *Soft computing* 9, 1 (2005), 3–12.

[4] Jun Zhang, Zhi-hui Zhan, Ying Lin, Ni Chen, Yue-jiao Gong, Jing-hui Zhong, Henry SH Chung, Yun Li, and Yu-hui Shi. 2011. Evolutionary computation meets machine learning: A survey. *IEEE Computational Intelligence Magazine* 6, 4 (2011), 68–75.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html