Application of Estimation of Distribution Algorithm for Feature Selection

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

ABSTRACT

Feature selection is a machine learning concept that entails selecting relevant features while eliminating irrelevant and redundant features. This process helps to speed up learning. In this paper, an Estimation of Distribution Algorithm (EDA) is applied to a feature selection problem originating from a legal business. The EDA was able to generate a realistic solution to the real-world problem.

CCS CONCEPTS

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

KEYWORDS

Feature Selection, Support Vector Machine, Estimation of Distribution Algorithm

ACM Reference format:
Mayowa Ayodele. 2019. Application of Estimation of Distribution Algorithm for Feature Selection. 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/10.1145/3319619.3326771

1 BACKGROUND

In recent years, feature selection has been of research interest in the computational intelligence community. It is an important preprocessing method in supervised learning. As many businesses are now collecting more and more data, feature selection is becoming increasingly important. It helps to speed up learning, improves the performance of machine learning algorithms and the quality of models [5].

In [6], feature selection approaches are categorised into *filters*, *wrappers* and *embedded* methods. Filters mostly differ from other methods because they use criteria that does not involve any machine learning. They are therefore used as a pre-processing step. Common examples are Pearson's correlation coefficient and information gain. Wrappers use machine learning models to evaluate the relevance of features, selection of features can be done using search algorithms like greedy search algorithms. Embedded models however perform selection as part of the construction of the model e.g L1 or L2 regularisation. Embedded methods attempt to reduce the computational effort required by wrappers to reclassify several feature subsets [4].

While Evolutionary Algorithms (EA) can be considered wrapper methods, they are reputable for significantly reducing the search space of solutions. When EAs are applied for feature selection, the selected machine learning model is used to evaluate each solution generated by the EA. A survey of EAs applied for feature selection such as Ant Colony Optimisation, Particle Swarm Optimisation and Genetic Algorithm is presented in [1].

Although there is no one algorithm that can be considered the best at solving all problems, Estimation of Distribution Algorithms (EDAs) have recorded competitive performances on binary problems [7]. Also EDAs can converge much quicker to good solutions than other EAs like the GA on problems they are suited for [2].

In this paper, an EDA is applied for feature selection where a fitness penalty approach is proposed to bias solutions towards less number of features. The problem considered in this paper originates from a legal business.

The rest of this paper is structured as follows. In Section 2, the feature selection problem and the business it originates from are described. Sections 3 and 4 respectively describes the solution approach and the results. Conclusions are presented in Section 5

2 PROBLEM

The problem presented in this paper is that of predicting the likely method of settlement for a claim. This problem originates from a legal business that represents insurance companies. It is important for the business to particularly identify matters that are likely going to be lost or won at trial. This will help the business make decisions regarding how to handle a matter. Other relevant methods of settlement are *Negotiation*, *Part 36 Offer* and *Withdrawn*. When a claim is settled by *Negotiation*, the business negotiates with the opponent claimant solicitors to get the best deal for its insurer clients. *Part 36 Offer* is used to describe an offer made as a tactical step designed to convince the opponent claimant solicitors to settle the claim early. The difference between *Part 36 Offer* and *Negotiation* is that there are more regulatory restrictions on *Part 36 Offers* than *Negotiation*. However, a matter may also get *Withdrawn*, this can happen when there is insufficient evidence by the opponent Claimant Solicitor. In summary, the methods of settlement considered in this paper are *Discontinued*, *Lost at Trial*, *Negotiation*, *Part 36 Offer* and *Won at Trial*.

2.1 Problem Features

The business captures several features when handling a matter. The features explored in this paper are presented in Table 1

GECCO '19, Prague, Czech Republic
 978-1-4503-6748-6/19/07…$15.00
DOI: https://doi.org/10.1145/3319619.3326771

Table 1: Features

Features	Description
Damages Claimed	How much has been claimed ?
Claimant Solicitor	Who are opponent claimant solicitors?
Client	Which client are we acting for?
Court	What is the location of the court?
Person Injury?	Was there any personal injury?
Injury Type	What is the category of the injury sustained?
Prospect Grading	How complex is this matter on a scale of 1-10?
Reason for Instruction	Why were we instructed?
Branch	Which branch was instructed?
Team	Which team handled the claim?

3 SOLUTION APPROACH

The classification algorithm used in this study is the well-known Support Vector Machine (SVM). The SVM is used because it is one of the most stable classifiers [3].

Alg. 1 describes the EDA used in this paper.

Algorithm 1 EDA for Feature Selection

1: Initialise ts, ps, gen and $M_{prob}(0)$
2: **for** $g = 1$ to gen **do**
3: Set $P = \emptyset$
4: **for** $i = 1$ to ps **do**
5: Generate ind by sampling $M_{prob}(g-1)$
6: Assign fitness to ind
7: Add ind to P
8: **end for**
9: Select best $ts < ps$ solutions to form S
10: Generate $M_{prob}(g)$ using S
11: Set $P_{new} = P$
12: **end for**
13: **return** best solution b in P

In 1, ts, ps and gen respectively denote truncation size, population size and number of generations. At each generation, a new probabilistic model $M_{prob}(g)$ is created and used to generate new solutions in the following generation. A new population P_{new} is populated at each generation and completely replaces that of the previous generation P. The best solution b is returned at the end of the run.

To generate the fitness, the Support Vector Machine (SVM) is applied to the selected features. The well-known 10-fold cross validation method is applied. The classification accuracy is then used to calculate the fitness. The fitness function is presented in 1.

$$f = accuracy - (n_{features}/1000) \quad (1)$$

Although the accuracy can be sufficient as the fitness function, a small penalty is applied so that the algorithm biases its search towards the lowest number of features $n_{features}$ required to get the optimal accuracy.

3.1 EDA parameters

The length of the solution is set to 10 which is the number of features considered in this paper while ts, ps and gen are respectively set to 5, 20 and 5. A total of 100 fitness evaluations has been used in this paper. A set of 3935 real-world examples was used to train the SVM.

4 RESULTS

In this section, results produced by the EDA are validated using the well-known feature importance method. Although the EDA was executed multiple times, the average cross-validation score produced was always 0.74. The EDA selected **Client** as the only feature for predicting method of settlement.

In Table 2, the feature importance values have been generated using the Extra Tree Classifier [1]. The feature importance attribute of this classifier computes the relative importance of all the features.

The choice of the EDA is comparable to the feature importance method as **Client** has a much higher score than any of the other features.

Table 2: Features

Rank	Feature	Importance Value
1	Client	0.481
2	Damages Claimed	0.084
3	Claimant Solicitor	0.080
4	Court	0.080
5	Prospect Grading	0.077
6	Team	0.059
7	Injury Type	0.056
8	Reason for Instruction	0.051
9	Branch	0.025
10	Personal Injury?	0.007

5 CONCLUSION

In this paper, an EDA has been applied for feature selection in a real-world problem which originates from a legal business. Although the problem considered is a simple one, the approach is applicable for a larger feature space. The EDA not only helps to select the feature with the highest importance but also selects the optimal number of features.

REFERENCES

[1] A Sheik Abdullah, C Ramya, V Priyadharsini, C Reshma, and S Selvakumar. 2017. A survey on evolutionary techniques for feature selection. In *Emerging Devices and Smart Systems (ICEDSS), 2017 Conference on.* IEEE, 58–62.
[2] Mayowa Ayodele. 2018. Effective and efficient estimation of distribution algorithms for permutation and scheduling problems. (2018).
[3] Ioan Buciu, Constantine Kotropoulos, and Ioannis Pitas. 2006. Demonstrating the stability of support vector machines for classification. *Signal Processing* 86, 9 (2006), 2364–2380.
[4] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40, 1 (2014), 16–28.
[5] Augusto Destrero, Sofia Mosci, Christine De Mol, Alessandro Verri, and Francesca Odone. 2009. Feature selection for high-dimensional data. *Computational management science* 6, 1 (2009), 25–40.
[6] Isabelle Guyon and André Elisseeff. 2006. *Feature Extraction: Foundations and Applications.* Springer Berlin Heidelberg.
[7] Mark Hauschild and Martin Pelikan. 2011. An introduction and survey of estimation of distribution algorithms. *Swarm and evolutionary computation* 1, 3 (2011), 111–128.

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