Genetic Algorithm Based Attribute Selection for a Credit Risk Classification Problem

Vaanee Nagpal

Research School of Computer Science, Australian National University
Canberra ACT 0200, AUSTRALIA
U5734316@anu.edu.au

Abstract. This report proposes a Genetic algorithm for feature selection in combination with one-hot encoding technique and logistic algorithm over the German Statlog Credit dataset. Comparative analysis was conducted over the mentioned dataset to observe the influence on computational accuracy before and after performing GA-based attribute selection. The prediction accuracies were calculated over the population of 65 individuals in 11 generations. The results were graphed and interpreted to display that the feature selection used in combination with logistic regression gave higher computational accuracies for both validation and test data. The results were also compared to a paper written by Paul O' Dea, Josephine Griffith and Colm O' Riordan about the feature selection using information theory technique in combination with neyral networks which also used Statlog German Credit dataset. It was revealed that the approach defined in this paper provided better test accuracy.
Keywords: Genetic Algorithm, Attribute Selection, One-hot Encoding, Logistic Regression, Statlog German Credit dataset

1 Introduction

Attribute selection can be defined as a process of selecting optimal subset(s) of attributes for training a machine model. In other words, attribute selection is used to reduce the dimensionality of the dataset to increase the execution accuracy. In this research, a publicly available dataset of Statlog German Credit is used which is located in UCI Machine Learning Repository (Dheeru, 2017). The dataset is related to the real world problem of classifying whether the credit risk associated with a person is good or bad. It is advantageous to perform research on such a dataset as in-depth data analysis for it can be used by banking industry. Moreover, optimal subset of attributes can be used to devise other methods to determine the credit risk.

An investigation will be carried to determine the prediction accuracy of the credit risk by applying one-hot encoding technique, genetic algorithm (GA) based attribute selection and logistic regression algorithm on the Statlog German Credit Dataset. The one-hot encoding technique will be used to transform the categorical attributes into a format that can be provided to our classification algorithm to execute efficiently which will be followed by attribute selection using GA (Babatunde, 2014). Genetic Algorithms are considered to be an efficient way to select features as user/coder can change functional configuration to enhance the outcome. We will implement the GA based attribute selection and logistic regression to determine the optimal subset of attributes that contribute towards increase in the accuracy of the predictive model.

The prediction accuracy calculated using GA based attribute selection will be compared with the baseline accuracy i.e. where no attribute is eliminated. This comparison will be utilised to analyse the benefits or lacks of benefits of implementing attribute selection using genetic algorithm on a classification problem.

2 The German Credit Data Set

The Statlog German Credit dataset (Hofmann, 1994) consist of information of 1000 loan candidates. Each candidate for the loan is described using 20 discrete attributes such as credit amount, duration, employment history etc. Out of those 20 attributes, 13 attributes are categorical and 7 attributes are numerical. These attributes are used as a factor for classifying that whether the credit risk is good or bad (goal). The original data file is modified by adding labels for each attribute to enhance the readability of data file.

3 Method

To analyse the prediction accuracy before and after attribute selection, the python library, Pytorch was be utilised. Pytorch calculated the prediction accuracy of both, before and after feature selection, has been performed by using combination of one-hot encoding, genetic algorithm and logistic regression over population of 65 individuals and 11 generations.

Prior to calculating the accuracy, the Statlog German Credit Dataset (Hofmann, 1994) was randomly split into training, testing and validation datasets using train_test_split function imported from sklearn.model_selection library. This provides an opportunity for the network to learn on a certain set of data, and once trained, validate or regularise itself on an unbiased data points. Successively, the network can test its gained knowledge upon the unseen data known as test data. In addition to this, randomly splitting the data helps in avoiding the presence of too many instances from the same target class. We performed 80:20 split on the whole data (1000 instances) where training data consisted of 800 instances and validation data contained 200 instances. The training data (800 instances) was further divided into 80:20 ratio to obtain the testing data i.e. the training data consisted of 640 instances and testing data consisted of 160 instances. Moreover, we encoded the output classes into binary form(0 = Good, 1 = Bad).

The underlying idea behind the application of genetic algorithm in this classification problem is to take the population of individuals and randomly select the individuals and determine the strongest among them all. The strength of each individual is computed using the predefined fitness function. The fittest individuals are further used to create new generation of off springs which then compete with the individuals in the old generation to replace them in the next generation. In conclusion we iterate on the population to create population with fitter individuals by utilizing operators such as selection, crossover and mutation (Joshi, 2017).

In order to implement the genetic algorithm over the above mentioned dataset following elements were taken into account-:

1. Individual Representation (binary)
2. Fitness Function
3. Selection: Demonstrates degree at which individuals would take part in the next generation.
4. Variation: This includes crossover and mutation.
5. Stopping Criteria: Determines when the algorithm should terminate.

Fig 1.0 represents the working of the genetic algorithm.

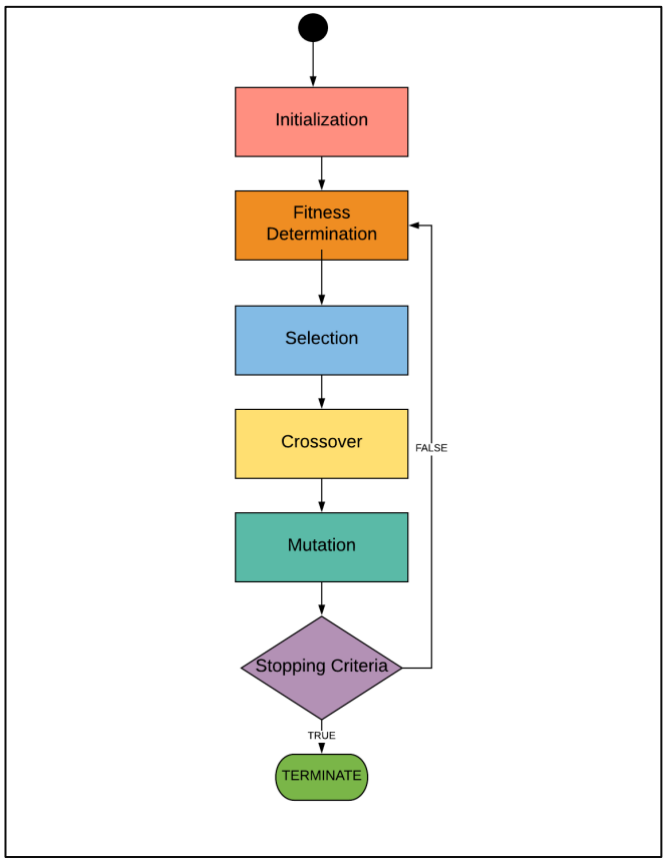

Fig 1.0: Diagram representing the working of a Genetic Algorithm

3.1 Advantages and Disadvantages of Genetic Algorithm

One of the advantages of using the genetic algorithm is that there is a low probability to get stuck at local optimum as we always store the set of potential individuals out of which one is the best. However, one of the disadvantages could be the randomness of the genetic algorithm. Genetic algorithms rely on the random sampling of individuals for iterating. In other words, the process is not deterministic i.e. different solution are obtained each time when the same algorithm is run. Therefore, every time we will run the algorithm we will have different subset of optimal features i.e. we would never obtain a fixed set of features which we can use to conduct further research on the above-mentioned dataset (Joshi, 2017)

.

4 Results and Discussion

4.1 Research Outcomes

The network was trained using the GA based attribute selection and logistic regression on Statlog German Credit Dataset. In order to set the usefulness of the above-mentioned approach we present the comparative analysis of the predicted accuracy with and without using the GA based attribute selection and logistic regression.

Initially 20 attributes were encoded into string of 0 and 1. These strings of zero and ones were used to create individuals which were then put into the population set. We calculated the fitness

of the individual and applied probabilistic operators such as mutation, crossover, selection. The fittest individuals from each generation were stored in a separate list. In conclusion this happened over the 11 generations where best individuals from each generation mated with each other to provide us the optimal subset of individuals(attributes).

We obtained the accuracy with all the features by simply applying Logistic Regression algorithm. The baseline accuracy for the test data and validation data was 73.75% and 78%.

After implementing the genetic algorithm, we obtained 10 subset features with the accuracy of 80% over the validation dataset and 9 subset features with the accuracy of 78.75% over the test data. Fig 1.1 shows the results over the validation dataset and Fig 1.2, shows the results over the test dataset

```
---Feature Subset From Validation Data---

Percentile:                    0.08438818565400844
Validation Accuracy:           0.81
Individual:     [0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1]
Number Features In Subset:     6
Feature Subset: ['stat_exist_CA', 'duration', 'credit_history', 'inst_rate_per', 'age', 'have_tele']
```

Fig 1.1: Represents predicted accuracy, fit individual, number of attributes in the subset , name of the attributes in the subset over validation data

```
---Feature Subset From Test Data---

Percentile:                    1.0
Test Accuracy:          0.79375
Individual:     [1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1]
Number Features In Subset:     7
Feature Subset: ['o_debtors', 'property_type', 'age', 'house_stat', 'no_credit_card', 'job_stat', 'p_maint.']
```

Fig 1.2: Represents predicted accuracy, fit individual, number of attributes in the subset , name of the attributes in the subset over test data

It can be observed that the removal of irrelevant features improves the computational accuracy over both validation and test dataset. The validation accuracy with 10 features is improved by 3% whereas the testing accuracy with 6 attributes has enhanced by approximately 6%. The 'Percentile' in Fig 1.1, defines the classification percentile of the best individual in the list over the validation data which is 0.0844. The 'Percentile' in Fig 1.2, defines the classification percentile of the best individual in the list over the test data which is 1.0. It is noteworthy that the distribution of the dataset also affects the predicted accuracy. The split proportions (training = 640 instances, testing = 160 instances and validation = 200 instances) can be considered as a good split.

4.2 Comparison of the result outcomes with the published paper

The study conducted by Paul O' Dea, Josephine Griffith and Colm O' Riordan in [2001] (Paul O' Dea, 2001) investigated the effects on predicted accuracy using feature selection in combination with Backpropagation algorithms (Neural Networks). They used the technique from information theory to select the attributes and then used Back Propagation Algorithm to train the neural networks. This study achieved the test accuracy of 74.25% with selecting 7 attributes. Furthermore, it was mentioned in the paper that a lot of time is required to train the network using their approach. In contrast we have used GA based attribute selection in combination with one hot encoding and Logistic Regression. Using our approach the time taken to train the network in comparatively small as well as our approach obtains the better test accuracy of 79.4%. Moreover, in Back Propagation algorithm there is always high probability to get stuck at local optimum.

5 Conclusion and Future Work

From the research conducted in this paper it was found that there was an incremental effect to computational accuracy by eliminating the irrelevant features using Genetic Algorithm based attribute selection upon Statlog German Credit Dataset. Better accuracies were achieved for both validation and test dataset by selecting attributes using genetic algorithm in combination with Logistic Regression.

It can be seen from Fig 1.3, that the testing accuracy obtained poor accuracy when we initially implemented our approach on the test data. However, the accuracy improved as the features were removed. The test accuracy over the selected features was significantly better than the test accuracy with all the features present.

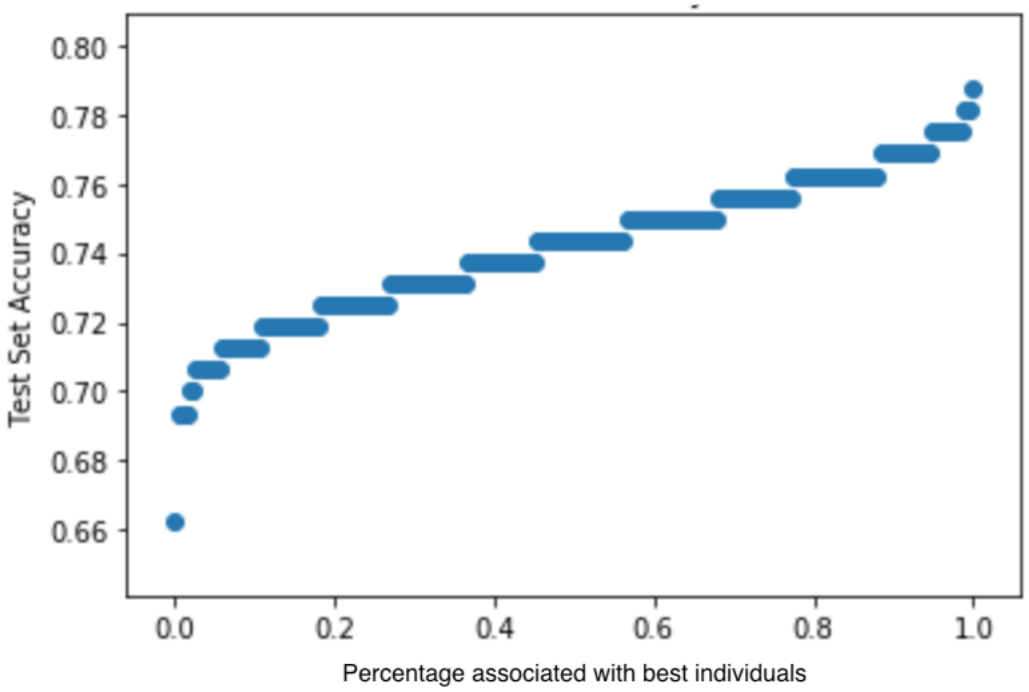

Fig 1.3: Graphical Representation of Test Accuracy(y-axis) against Percentage associated with best individuals in the population

From reviewing the external research paper by Paul O' Dea, Josephine Griffith and Colm O' Riordan in [2001] (Paul O' Dea, 2001) on the combination of feature selection using information theory in combination with neural networks it has shown room for further research by combining the certain aspects of both the approaches. Although our research provides better prediction accuracy, we always get different set of optimal features. Whereas, the computational accuracy is low as compared to ours, their approach suggests fixed features. Therefore, an in-depth comparison analysis between this research paper's result and Paul O' Dea, Josephine Griffith and Colm O' Riordan in [2001] research paper's outcome can be used to conduct further experiments and test. Example investigations could use one-hot encoding instead of thermometer encoding to encode the feature attributes and then apply combination of attribute selection using information theory and logistic regression.

Another suggestion for this research would be consider alternate datasets and combination of discrete techniques from both the papers. Moreover, this research can be extended using information theory for feature selections in combination with ID3 and Fuzzy Logic algorithms. However, this might not be approach to be used for the large and dense datasets.

6 References

- Hofmann, P. D. H., 1994. *UCI Machine Learning.* [Online]
  Available at: http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29
- Dheeru, D. a. K. T. E., 2017. *UCI Machine Learning Repository.* [Online]
  Available at: http://archive.ics.uci.edu/ml]
- Babatunde, O. &. A. L. &. L. J. &. D. D., 2014. A Genetic Algorithm-Based Feature Selection. *International Journal of Electronics Communication and Computer Engineering,* Issue 5, pp. 889-905.
- Joshi, P., 2017. *Artificial intelligence with python.* [Online]
  Available at: https://ebookcentral-proquest-com.virtual.anu.edu.au
- Paul O' Dea, J. G. C. O. R., 2001. *Combining Feature Selection and Neural Networks for Solving Classification Problems,* s.l.: s.n.