An Insight to Finding the Hidden Patterns in Genetic Algorithm and Data Mining

Abdulla Nabeel (S1700923)

1 INTRODUCTION

Machine learning and data mining algorithms are used to find the pattern in a data set that could be used to make people's life easier. Finding patterns manually by humans for a large scale of data is impossible. Therefore, we use these algorithms to find the patterns. A lot of algorithms are available to solve these kind of data mining problems such as pattern recognition problems. This report has been written on the attempts made to solve a set of classification problem using an evolutionary algorithm.

2 BACKGROUND RESEARCH

Data mining is the analysis of large scale data to extract hidden predictive information and making it useful. This is done by exploring, analyzing, reducing and reusing these massive amount of data to find the patterns in these bulk of data. Data mining can be used in any field. One of the major fields where data mining is used is in the medical field. Diagnosis of the patient record is used and analyzed to identify the best practices. Moreover, data mining is used by telecom companies, insurance companies and by banks to detect frauds. Furthermore, data mining can be used to identify criminals (David L.Olson, 2008).

For example, to find the patterns hidden in the data collected and recorded by banks, it is a necessity to employ an algorithm that has the ability to mine data and extract knowledge from it. Analyzing the bank transactions and using bio-inspired algorithms such as Artifical Immune Systems have been used to enhance the results of extracting the knowledge (Cunha & Castro, 2018). The use of Artifical Immune Systems has been increasing as it is capable of finding various solutions to vast problems. This way, the bank can use the extracted knowedge to improve their facilities and provide better services to the users.

Another application of data mining techniques is used in grocery stores or department stores. The stores offer loyalty cards to customers who wants to avail it. After availing, these members get offers and discounts which are not available for the regular customers. The use of these cards make it easier for the stores to gather a vast amount of data regarding the loyalty member's preferences. After analyzing those records, the stores can come up with sales and promotions targerted to their customers (Michael J.A. Berry, 2004).

Moreover, the field of education also makes use of data mining techniques. The main purpose of it is to improve the assessment of the students's learning processes and achievements by predicting information from the edicational databases. Educational data mining usually relies on techniques such as k-nearest neighbor, neural networks, decision trees, support vector machines and naive bayes. The data obtained can be used in making effective decisions in striving and improving the students performaces (Jalota & Agrawal, 2019).

Evolutionary computing has been used in data mining in the field of computer networking. Evolutionary programing take up representations that are more normal or usual for the tasks instead of relying on the singular and general representation such as in genetic algorithm. In cases where there aren't enough calculus for guiding recombination or when separating the search and evaluation spaces does not seem useful, evolutionary programming is usually preferred over genetic algorithms (Peter J. Angeline, January 1994)

There are numerous advantages of using evolutionary algorithms. One of the most important of such advantages is that evolutionary algorithms

works based on the manner of 'survival of the fittest'. It works by getting rid of any factors that are not suitable or appropriate but by using the knowledge in its original solution. This leads to the selection of the best solutions. Another advantage is that evolutionary algorithm uses natural evolutionary mechanisms to resolve the difficult problems. Moreover, even if evolutionary algorithm is used, other applications can be used along with it to find an optimal solution (Ming Xue, 2009).

3 EXPERIMENTATION

3.1 Setting up Algorithm

The purpose of this experimentation is to create a universal rule set that can correctly classify a given set of input variables and give the predicted output. This is done by generating a random population, same as the provided datasets. Then the fitness is calculated for the individuals after that, the selection is applied. The selection also differ by the data set selected by the user which is dynamically updated in the programme. The details of this will be provided later in this report with each dataset and the difference made to the programme to adjust with the dataset. After that, the selected individuals undergoes a crossover process and the individuals then is mutated and passed to the next generation. The program is dynamic so that the parameters can be changed or manipulated to see the difference which can be evaluated to find the optimal solution.

Genes
An individual is characterized by a set of parameters known as genes. Genes are used as parameters of an individual's characters. These characters are often represented in bits of binaries in the genes. These genes are represented as data which need to be classified. The main purpose of this report was to classify the given 3 datasets. Therefore, to achieve this, the data in the files should be represented as genes.
Example:
10001 1
In the above example each digit "1" or a "0"is represented as a gene.

Initial Population
Population is the set of individuals which consists of a set of chromosomes. To initiate the algorithm, a population should be generated which could be of any size. From few individuals to thousands. In this case, population is generated similar to the provided dataset with the inputs and outputs. '10000 1'. In this case the input is '10000' and the output is "1". The algorithm should be able to classify the inputs and give the result of '1'. While generating the random population a wildcard was introduced to the input population. In the dataset one and two, a wildcard of '2' was introduced. And for the dataset three, a wildcard of '0.5' is used.

Tournament Selection
To improve the populations overall fitness, the individuals with the highest fitness score should be selected and passed onto the next generation. In addition to that, those individuals with the less fitness score should be discarded from the population. Therefore to achieve this, tournament selection is used in the dataset one and two as the dataset is fewer when compared to the dataset three. For the few datasets the tournament selection works best. Tournament selection uses a strategy that is used to select the fittest candidate amongst the generation and is passed to the next generation.

Roulette Wheel Selection
For the dataset three the selection that was used is roulette wheel selection. In this selection, proportion of the wheel is assigned by the fitness value. Therefore the better the fitness value, better the chance of getting selected to the next generation. Moreover, there is a chance that an individual with the less fitness value be selected to the next generation. Although the probability of this selection is low, it is not zero. This helps the algorithm to avoid getting stuck on the local maxima.

Crossover
Crossover is used to mix-up the individuals of the population, intention of creating new individuals with higher fitness. In this function, two individual is selected and by using a random number generator to select the point of a gene from where the swap should be. This allows individuals to swap their heads and tails. A new individual with higher fitness is created by merging the head of high fitness score

and tail of high fitness score. Therefore this creates a much 'fitter' offspring which inherits the best traits from their parents.

Mutation

Mutation is a necessary function to be implemented as it adds randomness to the individual's genes. Without the mutation there is a high chance that the solutions we get would be already in our initial population. Moreover, the tournament selection discards all the weakest individual. As a result, this would prevent from reaching to the goal state which leaves the algorithm in a local maxima. Therefore, to prevent this a subtle changes at random to the individuals genome is added.

Evaluation

Evaluation is used to score the population to find the fittest solution. If the current solution score is greater than the solution score being compared, then it is updated as the fittest.

3.2 Dataset one

Data Representation

In the dataset one there are 32 rows of bit strings, '00001 1'each of 5 inputs and 1 output. If the input bit strings ('00001') are presented, the solution of "1" bit should be presented as a result. As mentioned above, an initial population is generated randomly with the bit strings '0', '1' and '2' while '2'being a wildcard.

In the fitness function, the randomly generated values are validated with the validation set that was in the initial dataset. If the validation set and the randomly generated set values match, the fitness score is increased. Moreover, if the randomly generated chromosome contains a wildcard of "2", then it is considered a match. For example if the validation set is '10001' and the randomly generated value is '20002' than it is considered as a match and the fitness score is increased.

Parameter Testing Dataset 01

To test the effect of parameter changes to the fitness over the population, the following is kept constant.

- *Size of population = 1000*
- *Number of iteration / generation = 1000*
- *Chromosome size = 32*

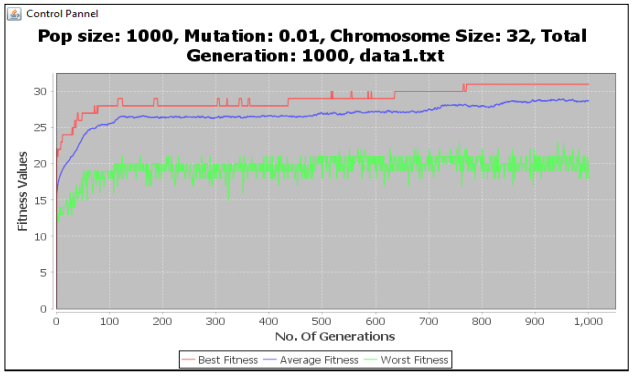

Figure:01

Figure 01 shows the effect of mutation. On the X-axis is the number of the generation and on the Y-axis is the fitness values. The graph shows the best fitness, average fitness and worst fitness per generations. The best fitness score achieved in the set was 31. Therefore the accuracy is about 96%.

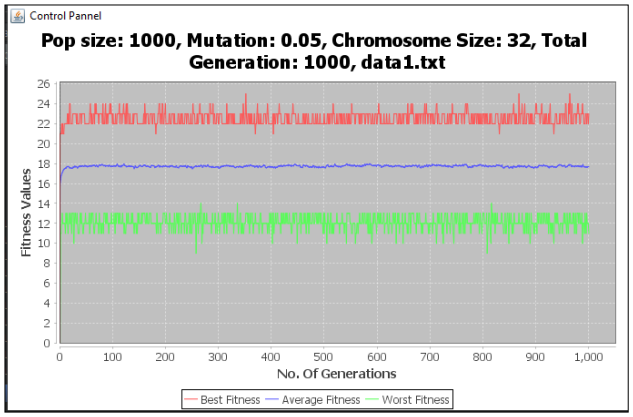

Figure: 02

Figure 02 shows the effect increase of mutation to the data set one. As seen in the diagram there is a significant change in both the diagrams when compared with the figure 01. In figure 02, there is much greater oscillation in the best fitness as the mutation rate increased. The figure 01 has steady best fitness after the 800th generation. However, the fitness also gradually decreased with the increase of the mutation. Figure 01 had 31 as the best fitness, however, figure 02 has 25 as the best fitness. The accuracy is about 75%. Therefore this proves that the individuals are stuck at local maxima. Moreover, the average fitness also decreased in figure 02 when compared with the figure 01.

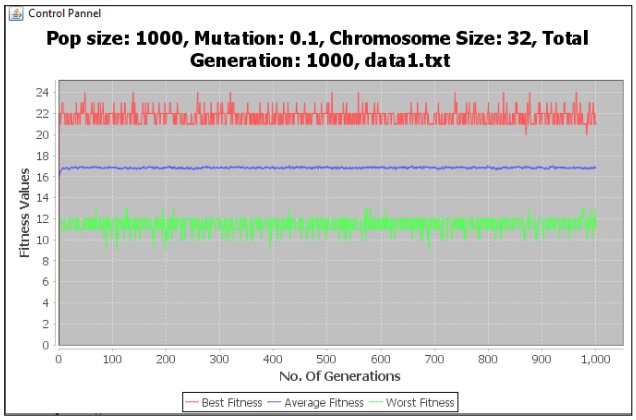

Figure: 03

Figure 03 shows the graph with the mutation rate 0.1. The significant change that is noticed in the graph is that the oscillation of the best fitness increased even more as the mutation rate increased. The accuracy also gradually decreased when compared with other two graphs. The accuracy is about 68%. Moreover, the average fitness also decreased with the increase in the mutation rate.
Therefore, these graphs prove that the mutation rate should be kept at a lower rate to achieve an optimum result. Furthermore, to get a steady fitness across the generations the mutation should be kept at a lower rate.

3.3 Data Set 2
Differences

The dataset one has five bit string as input and one bit as a result in a row, with a total of 32 rows. But the dataset two has six bit as input and one bit as result in a row and the file contains 64 rows. Apart from those differences, both datasets are pretty much the same in nature. Therefore, this change has to be made in the algorithm.
The user can select which dataset to run, and the programme will dynamically adjust the changes. Therefore, when the user selects the dataset 02, the length of the chromosome will be updated.

Parameter Testing Dataset 02
To test the effect of parameter changes to the population over the fitness, the following is kept constant.

- *Mutation = 0.02*
- *Number of iteration / generation = 1000*
- *Chromosome size = 32*

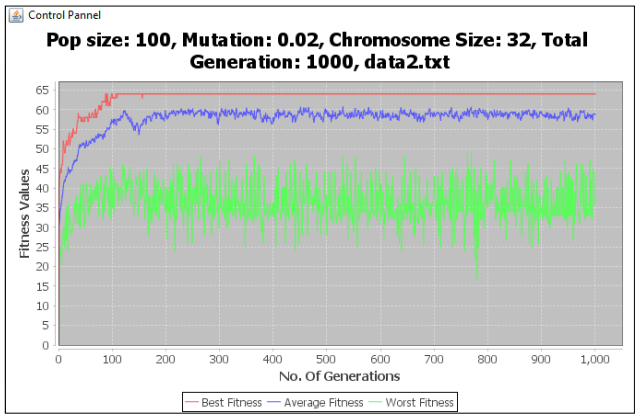

Figure: 04

Figure 04 shows the effect of population over the fitness. As seen in the above figure, the population is set to 100 and it takes 110 generations for the best fitness to reach to the global maxima.

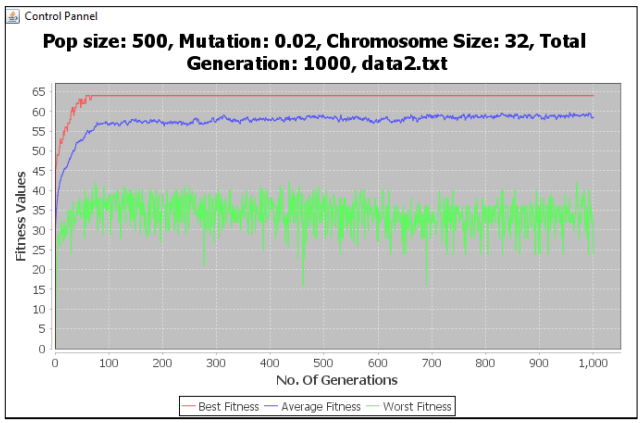

Figure: 05

Figure 05 shows the effect of population over the fitness. The population is set to 500 and the result is shown. The figure shows that to reach the global maxima it takes only about 60 generations. Therefore, this proves that the increasing population could help to reach the global maxima even faster.

3.4 Data Set 3
Data Representation

The main difference that was noticed was, unlike the dataset 1 and 2, the dataset 3 contains floating point numbers. Besides, the dataset contains 2000 rows. Major changes in the algorithm should be made to compensate this requirement.

Changes to Algorithm

The current function to populate gene populates binary values. Therefore, this function should be changed. Moreover, with this update, fitness function should also be changed.

The function which generates the population now generates floating point values. The updated fitness function now uses '0.5' to see if the randomly generated value matches the validation set values. This is done by evaluating if the values is less than or greater than or equal to the validation set and the randomly generated set, and a score is increased if the validation set value and the randomly generated value matches. For example if the validation set value is 0.75 and the randomly generated set value is 0.80, then it is considered as a match and fitness score is increased. But if the validation set is 0.20 and the randomly generated set value is 0.60 then, it is considered a not match and fitness score will not be updated.

Half of the dataset is used to train the algorithm and other half is used to test the algorithm.

Parameter Testing Dataset 03

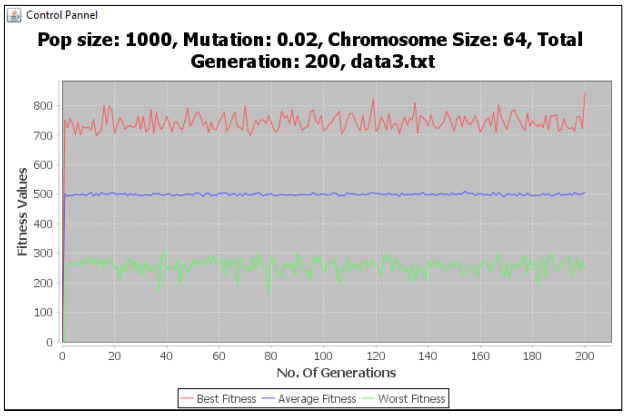

Figure: 06

Figure 06 shows the performance of the dataset 03 using roulette wheel selection. One of the significant change noticed was that the processing time was greater than dataset 1 and 2 as both the dataset used tournament selection, it was much faster. But the dataset 3 uses roulette wheel selection which greatly increases the diversity. Hence, it increases the processing time. However, the accuracy of dataset 3 is about 40%.

4.0 CONCLUSION

This paper highlights how the evolutionary algorithms are used in data mining, and how the data classified is used with examples. In addition to that, this paper includes experimentations of creating an evolutionary algorithm to classify 3 datasets. With the parameter changes the results are analysed. This experimentation proves that, to increase the fitness it is necessary to have mutation as it helps to avoid individual getting stuck on the local maxima. According to this experimentation the appropriate level is 0.01. Moreover, this experimentation proves that having a good population size helps to reach the maximum fitness much faster than having a less population size. In addition to that, roulette wheel selection increase the diversity. The accuracy of this algorithm can be further increased by neural network.

With this experimentation it can be concluded that Genetic Algorithm can be used to find the hidden patterns in the data that can be used to classify these datasets. Therefore, it proves that GA is a powerful tool that with the enough data and proper perimeter changes it is capable to find an optimal solution.

REFERENCES

Cunha, D. S. d. & Castro, L. N. d., 2018. *Evolutionary and Immune Algorithms Applied to Association Rule Mining in Static and Stream Data.* [Online]
Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8477978
[Accessed 2020 January 4].

David L.Olson, D. D., 2008. *Advanced Data Mining Techniques.* 1st ed. Lincoln: Verlag Berlin Heidelberg.

Jalota, C. & Agrawal, R., 2019. *Analysis of Educational Data Mining using Classification.* [Online]
Available at: https://ieeexplore.ieee.org/document/8862214
[Accessed 4 January 2020].

Michael J.A. Berry, G. S. L., 2004. *Data Mining Techniques For Marketing, Sales and Customer*

Relationship Management. 2nd ed. Indiana: Wiley Publishing Inc..
Ming Xue, C. Z., 2009. The Application of Data Mining In the Decision of Supermarket Extension and Businesses Expansion Based on Evolutionary Computation. *IEEE Computer Society*, p. 779.
Peter J. Angeline, G. M. S. a. J. B. P., January 1994 . An Evolutionary Algorithm that. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 5(1), p. 56.

Github:
https://github.com/Lyban/BioComputation_Assignment