Assessment 1.2: Feature Selection for Death Prediction by Heart Failure

Anas Javed, ID: 22118871 [*1]

Word Count: 2200

Abstract

Predicting survival of heart failure patients is crucial for hospitals and medical practitioners. Knowledge of key values causing death can play important part in saving patient's life. The objective of this report is to find the most important features causing death by heart failure. In this report, a genetic algorithm with three different variations is used to select the best features from "Heart Failure Clinical Record Dataset". Three different crossover techniques used are: Single point crossover, uniform crossover and blend crossover. Crossover and mutation probabilities has been demonstrated in this report. The algorithms were designed to maximize the accuracy while using the minimum number of features. The results show that single point crossover with the large population size performs better than the other two methods. Key features selected by the algorithm were ejection fraction, high blood pressure and platelets. In this report we analysed that by carefully controlling the vital values, survival probability can surely increase.

1. Introduction

Cardiovascular diseases (CVD) are the leading cause of mortality worldwide. Every year, an estimated 17.9 million people die of these diseases, accounting for 31% of all fatalities worldwide (Henkel et al., 2008). CVD kills an estimated 3.8 million men and 3.4 million women each year. These figures are large enough to seek the attention of the scientists.

The phrase "cardiovascular disease(CVD)" refers to any ailment that affects the heart or blood vessels (Spear & Dismukes, 1994). There are several key factors playing their part in the CVD. The most important ones are high blood pressure, smoking, high blood cholesterol, diabetes, lack of activity, obesity and family history. Few major complications include angina, heart attack and heart failures. Angina is the restricted blood supply to the heart muscle causing chest discomfort. During heart attack, the blood supply to the heart muscle is suddenly cut off. Heart failure is one of the most fatal issue. It occurs when the heart is unable to efficiently pump blood around the body (Amma, 2012). We used population based genetic algorithm with three different types of crossover operators to predict the death rate due to heart failure. It incorporates maximizing the accuracy of prediction with less number of features. After feature selection, decision tree classifier is used to predict the death rate. It shall definitely help in increasing the life expectancy. The remaining part of the report consists of section 2, highlighting the previous work. Section 3 is giving details about the data set, objective function and candidate solution. Section 4 is describing the methods being used in this research. Section 5 details the configuration parameters and working of the three methods used. Section **??** and 7 explains the results obtained from the three methods.

[1]M.Sc. Big Data Analytics, School of Computing and Digital Technology, Birmingham City University, UK. Correspondence to: Anas Javed <anas.javed@mail.bcu.ac.uk>.

Modern Optimisation Report, CMP 7213 - Modern Optimisation.

2. Literature Review

In the previous studies for the prediction of death in heart failure, different meta-heuristic algorithms have been used. Most common algorithms are Genetic Algorithm (GA) with different crossover techniques like single point crossover, multiple point crossover, uniform crossover, arithmetic crossover with different selection criteria like Roulette wheel and tournament selection has been used for death prediction (Mathew, 2012; Umbarkar & Sheth, 2015). In a paper, Whitley (1994) used Genetic Algorithm with Single Point(SP) crossover to select the features with 300 population size. Accuracy results deduced using SP crossover were 82%, those were quite significant, keeping in the view, they used the minimum features to reduce the computational cost. The usage of these parameters are discussed in the Experimental Setup section 5. Multiple studies with Support Vector Machine (SVM) (Tao et al., 2018), Decision Tree and Naive Bayes classification are also reviewed to understand the problem in detail (Tanaka et al., 1995). In the scope of this paper, we are using decision tree for classification and

GA for feature selection.

3. Problem Instance

In the scope of this paper, we shall be discussing the cardiovascular diseases and its mortality rate in human beings. What are the key factors that cause the heart failure and how do they lead to the increase in mortality rate? We may use a variety of variables to predict mortality, but our objective is to employ the most significant or crucial elements to forecast death events. It can be done by feature selection using genetic algorithm (Panda et al., 2019).

3.1. Data Set

Dataset is taken from kaggle. It is an open source online platform where data set for many of the data science problems are available. Data is provided by Davide Chicco and Giuseppe Jurman in "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020)." It is named as "Heart Failure Clinical Record Dataset".

The dataset contains entries of 300 patients with the following features: Age, Anaemia, Creatinine Phosphokinase, Diabetes, Ejection Fraction, High Blood Pressure, Platelets, Serum Creatinine, Serum Sodium, Sex, Smoking and Time.

3.2. Objective Function

Our objective is to minimize the number of feature while maximizing the accuracy.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Where
TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

$Objective function = Accuracy * (1 - \frac{Features used}{Total Features})$

The goal is to maximize the objective function. Hence, making it a maximization problem.

3.3. Candidate Solution

Our candidate solution is the vector of one and zeros. It looks like: (1,0,1,0,0,0,1,0,1,0,1,1)

where the 1's represents the features used for the prediction and 0's represents the features that are not being used for prediction.

Fitness value is calculated by adding all the 1's representing features used and dividing the sum by the total number of features.

3.4. Modern Optimization Methods

This article will be using the following three optimization methods to compute fitness value and will find the method that performs better for this optimization problem.

- Genetic Algorithm (Single Point Crossover)
- Genetic Algorithm (Uniform Crossover)
- Genetic Algorithm (Blend Crossover)

4. Methodology

This section discusses the three strategies that we used in the feature selection.

Genetic Algorithm

Population based genetic algorithm (GA) is the most common method used widely for feature selection. GA is based on Darwin's "survival of fittest" theory (Whitley, 1994). In GA, solution set contains most suited set of features at the point when we stop our computation. Other poor quality solutions may be ruled out during crossover and mutation. In GA, crossover is a very common phenomena while mutation is a rare one (Mathew, 2012).

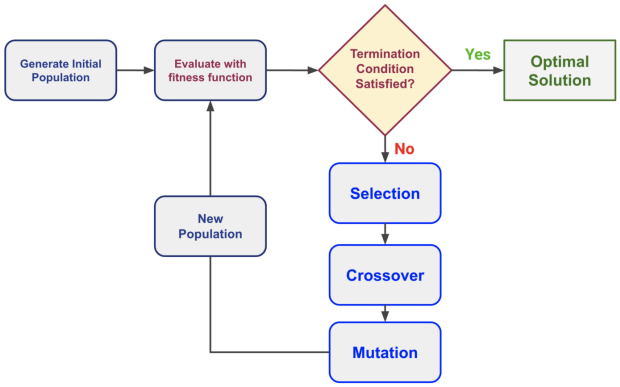

Figure 1. Complete workflow of Genetic Algorithm Explained (Mathew, 2012)

4.1. Genetic Algorithm (Single Point Crossover)

In single point crossover, a random crossover point is selected and tail of the parents swap to produce the new offspring (De Jong & Spears, 1992). Our GA with single point crossover starts with randomly selecting two parents' population. Single point crossover and mutation is applied to produce offspring. Then, the offspring are put back to the population. After that fitness score is assigned to the offspring after evaluation. If it's the best solution so far will

note that. We shall repeat the crossover and mutation unless we meet the stopping or termination criteria. We have our evaluation criteria with us and assign every member of the population a score and a fitness value (Umbarkar & Sheth, 2015).

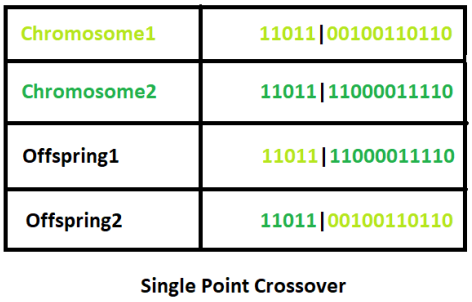

Figure 2. Genetic Algorithm with Single Point Crossover (Umbarkar & Sheth, 2015).

4.2. Genetic Algorithm (Uniform Crossover)

In uniform crossover, each bit is selected randomly from one of the parent's chromosome. Our GA starts randomly by selecting two parents' population. We apply uniform crossover and mutation to produce offspring. Then, the offspring is put back to the population. After that, the offspring is evaluated and fitness score is assigned to the offspring. If it's the best solution so far, will note that. GA repeats the crossover and mutation unless the stopping or termination criteria is reached. We have our evaluation criteria with us and assign every member of the population a score and a fitness value (Hu & Di Paolo, 2009).

4.3. Genetic Algorithm (Blend Crossover)

In blend crossover, the blend crossover randomly selects a child in the range $[x1 - \alpha(x2 - x1), x2 + \alpha(x2 - x1)]$ (Reddy et al., 2020). the parent's chromosome. Our GA starts randomly by selecting two parents' population. We applied blend crossover to produce offspring. Then, the offspring are put back to the population. After that, the offspring are evaluated and fitness score is assigned to the offspring. If it is the best solution so far, will note that. The GA repeats the crossover and mutation unless it met the stopping or termination criteria. We have our evaluation criteria with us and assign every member of the population a score and a fitness value (Reddy et al., 2020).

5. Experimental Setup

The parameters of genetic algorithms used to conduct the experiments are demonstrated in the table 1,table 2 and table 3.

5.1. Genetic Algorithm (Single Point Crossover)

The selection parameter for the single point crossover has been referenced from the articles cited here: (Hasançebi & Erbatur, 2000) (Srinivas & Patnaik, 1994). Due to high computational cost, number of generations is set to 20. Weighted Roulette wheel is used to give a solution the better opportunity to get selected. Refer to Table 1 for detailed parameters.

Parameter	Value
Population Size	200
Maximum Generations	20
Type	Binary
Operator Probabilities	Crossover: 0.9
Operator Probabilities	Mutation: 0.1
Crossover Operator	GA SP Crossover
Selection Operator	Weighted Roulette Wheel
Method	Classification
Classifier	Decision Tree
Dataset	Heart Failure Clinical Record Dataset

Table 1. Configuration parameters for Genetic Algorithm (Single Point Crossover)

5.2. Genetic Algorithm (Uniform Crossover)

The selection parameter for the uniform crossover has been referenced from previous studies (Syswerda et al., 1989). There are few changes in the parameters to improve the accuracy and fitness value. Refer to Table 2 for detailed parameters.

5.3. Genetic Algorithm (Blend Crossover)

Third method used in this feature selection problem is GA with blend crossover. We have changed the population size, operator probabilities to note some significant difference in the fitness values. Other parameters has been referenced from the previous research article (Takahashi & Kita, 2001). We have changed the parameters multiple times to find the better results, these are the best parameters we found so far. Refer to Table 3 for detailed parameters.

Parameter	Value
Population Size	300
Maximum Generations	20
Type	Binary
Operator Probabilities	Crossover: 0.85
Operator Probabilities	Mutation: 0.15
Crossover Operator	GA Uniform Crossover
Selection Operator	Weighted Roulette Wheel
Method	Classification
Classifier	Decision Tree
Dataset	Heart Failure Clinical Record Dataset

Table 2. Configuration parameters for Genetic Algorithm (Uniform Crossover)

Parameter	Value
Population Size	150
Maximum Generations	20
Type	Binary
Operator Probabilities	Crossover: 0.8
Operator Probabilities	Mutation: 0.2
Crossover Operator	GA Blx Crossover
Selection Operator	Weighted Roulette Wheel
Method	Classification
Classifier	Decision Tree
Dataset	Heart Failure Clinical Record Dataset

Table 3. Configuration parameters for Genetic Algorithm (Blend Crossover)

5.4. Classification

In this report, we have used decision tree for classification. Decision tree is a non-parametric supervised learning approach. The objective is to learn basic decision rules from data attributes to develop a model that predicts the value of a target variable. We use decision tree because it divides the data into small manageable parts. As we have labeled data set it is better to used supervised machine learning approach. Our split ratio for the model was 70:30, where 70 percent was training data and 30 percent was testing data.

6. Results

The experiments conducted using the parameters mentioned in the table 1, 2 and 3 shows that some features holds significant value while others do not play much part in death prediction in heart attack. It is clear from the results that most important features are anemia, diabetes, ejection fraction platelets and Smoking because these features has been selected multiple times in different experiments.

Feature	Single Point Crossover	Uniform Crossover	Blend Crossover
Age	0	0	0
Anaemia	1	0	1
Creatinine Phosphokinase	1	0	0
Diabetes	1	1	1
Ejection Fraction	0	1	1
High Blood Pressure	0		0
Platelets	0	1	1
Serum Creatinine	0	0	0
Serum Sodium	0	0	0
Sex	1	0	1
Smoking	1	1	0
Time	0	0	0

Table 4. Features selected for various GA's

After running the experiments under the careful observation and control conditions the following graph (See Figure 3) is formed. In the graph, X-axis represents generation number while y-axis represents it's corresponding fitness score. This fitness value is normalised between 0 and 1 to make it more meaningful, where 0 means minimum and 1 represents maximum.

Our objective was to maximize the accuracy while using the minimum number of feature and eventually maximize the fitness score. In the experiments performed we can see that in the table 5 Genetic Algorithm with uniform crossover presented the best fitness score while one point crossover also gave almost similar results. Performance of blend crossover was not very impressive as compared to other two methods.

Experiment	Fitness Score
Single Point Crossover Size	0.687
Uniform Crossover	0.689
Blend Crossover	0.684

Table 5. Fitness Score of Different Experiments

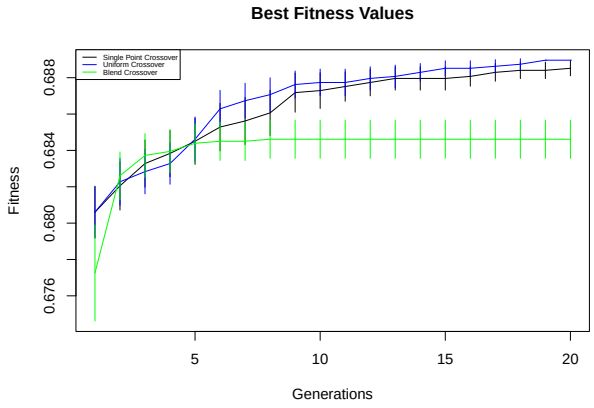

Figure 3. Best fitness of various methods on the Feature Selection problem against the number of generations (X-axis). GA with uniform crossover is clearly the best method.

7. Discussion

The goal of this report was to find the optimal set of features for survival prediction in heart failure. Experiments were designed to depict the performance of different crossover methods. Results shows that performance of uniform crossover method is much better than the blend crossover. In the chosen methods we can see that fitness score improves with every generation in all three methods. Trend lines in the graphs shows the increase in the fitness value with every generation and it's error bars.

The best result occurred in uniform crossover method in which only four features diabetes, ejection fraction, platelets and smoking were selected by the GA. This tells the story about the significance of these features. In an alarming condition like heart failure having a keen eye on blood sugar level, platelets count and ejection fraction can play a predominant part in saving life. Controlling these parameters by some artificial means can impact patient's health in positive manner.

In the above mentioned experiments, it's clear that with the increase in population, single point crossover performed the best among the other two methods. Figure indicates that there is a slight change in the accuracy after certain number of generations which means that once GA learns the model it converges to its best.

8. Conclusion and Future Work

This research was conducted to explore the key features playing remarkable role in life and death prediction. We have used GA with three different crossover operators and multiple population sizes. It's clear from the results that changing population sizes, crossover operators and selection criteria have consequential importance as they impact substantially on model's performance. Results shown in this report says that GA with uniform crossover and large population size performed better than other two methods. In this report, we highlighted that by controlling ejection fraction, high blood pressure, serum sodium and platelets count patient's survival rate can be increased largely. These results will surely help health practitioners and researchers to predict the survival probability after heart failure. After carefully analyzing the results, we found that there is still some room for improvement. As we have used base variants of GAs and they have performed to limited extent. In Future, we can try devising hybrid GA techniques to increase accuracy and fitness score (Doppala et al., 2021). It is also observed that acquiring larger dataset can also elevate accuracy to large extent while using less features. There are a lot of other datasets available on web. Repeating these experiments on different datasets may result in exploring better results.(Hasançebi & Erbatur, 2000)

References

Amma, N. B. Cardiovascular disease prediction system using genetic algorithm and neural network. In *2012 International Conference on Computing, Communication and Applications*, pp. 1–5. IEEE, 2012.

De Jong, K. A. and Spears, W. M. A formal analysis of the role of multi-point crossover in genetic algorithms. *Annals of mathematics and Artificial intelligence*, 5(1): 1–26, 1992.

Doppala, B. P., Bhattacharyya, D., Chakkravarthy, M., and Kim, T.-h. A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset. *Distributed and Parallel Databases*, pp. 1–20, 2021.

Doust, J. A., Pietrzak, E., Dobson, A., and Glasziou, P. How well does b-type natriuretic peptide predict death and cardiac events in patients with heart failure: systematic review. *Bmj*, 330(7492):625, 2005.

Hasan, N. and Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. *Health and Technology*, 11(1):49–62, 2021.

Hasançebi, O. and Erbatur, F. Evaluation of crossover techniques in genetic algorithm based optimum structural design. *Computers & Structures*, 78(1-3):435–448, 2000.

Henkel, D. M., Redfield, M. M., Weston, S. A., Gerber, Y., and Roger, V. L. Death in heart failure: a community perspective. *Circulation: Heart Failure*, 1(2):91–97, 2008.

Hu, X.-B. and Di Paolo, E. An efficient genetic algorithm with uniform crossover for air traffic control. *Computers & Operations Research*, 36(1):245–259, 2009.

Mathew, T. V. Genetic algorithm. *Report submitted at IIT Bombay*, 2012.

Panda, D., Ray, R., Abdullah, A. A., and Dash, S. R. Predictive systems: Role of feature selection in prediction of heart disease. In *Journal of Physics: Conference Series*, volume 1372, pp. 012074. IOP Publishing, 2019.

Prakash, S., Sangeetha, K., and Ramkumar, N. An optimal criterion feature selection method for prediction and effective analysis of heart disease. *Cluster Computing*, 22 (5):11957–11963, 2019.

Reddy, G. T., Reddy, M., Lakshmanna, K., Rajput, D. S., Kaluri, R., and Srivastava, G. Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evolutionary Intelligence*, 13(2):185–196, 2020.

Reddy, N. S. C., Nee, S. S., Min, L. Z., and Ying, C. X. Classification and feature selection approaches by machine learning techniques: Heart disease prediction. *International Journal of Innovative Computing*, 9(1), 2019.

Spear, K. E. and Dismukes, J. P. *Synthetic diamond: emerging CVD science and technology*, volume 25. John Wiley & Sons, 1994.

Srinivas, M. and Patnaik, L. M. Genetic algorithms: A survey. *computer*, 27(6):17–26, 1994.

Syswerda, G. et al. Uniform crossover in genetic algorithms. In *ICGA*, volume 3, pp. 2–9, 1989.

Takahashi, M. and Kita, H. A crossover operator using independent component analysis for real-coded genetic algorithms. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546)*, volume 1, pp. 643–649. IEEE, 2001.

Tanaka, M., Watanabe, H., Furukawa, Y., and Tanino, T. Ga-based decision support system for multicriteria optimization. In *1995 IEEE international conference on systems, man and cybernetics. Intelligent systems for the 21st century*, volume 2, pp. 1556–1561. IEEE, 1995.

Tao, P., Sun, Z., and Sun, Z. An improved intrusion detection algorithm based on ga and svm. *Ieee Access*, 6: 13624–13631, 2018.

Umbarkar, A. J. and Sheth, P. D. Crossover operators in genetic algorithms: a review. *ICTACT journal on soft computing*, 6(1), 2015.

Whitley, D. A genetic algorithm tutorial. *Statistics and computing*, 4(2):65–85, 1994.