Feature Selection with Genetic Evolutionary Algorithm

***Abstract*—In real-world problems, a large number of features are needed when we need to apply machine learning algorithms. But, not all features are essential because many of them are unnecessary or even irrelevant, which can reduce the effectiveness of any algorithm. Feature selection (FS) in machine learning is an important task to reduce the dimensionality of the data and increase the effectiveness of the algorithm. Various methods have been applied to solve FS problems, where evolutionary algorithms (EA) have recently gained considerable attention. The aim of this work is to apply different machine learning methods, optimized by genetic algorithm (GA) and compare the best output. In this analysis, a GA combined with an artificial neural network (ANN) showed the best prediction accuracy. Without GA the prediction accuracy is lowest in all methods, with GA the accuracy was increased 76.00% to 76.50%, 75.25% to 76.00%, and 74.92% to 75.83% in ANN, SVM, and LR, respectively. We got the best recall of 95.39% in the LR model which is better for this model than 50% and the F1 score is highest on the LR model and all methods showed results above 50%. That means all models fit best to our data. Therefore, developing filter measures specifically according to the characteristics of an EC technique may significantly increase the efficiency and effectiveness.**

Keywords—feature selection, evolutionary algorithms, genetic algorithm, machine learning, classification

I. Introduction

Selecting a subset can used for reducing the level of data for increasing the efficiency of the classification algorithm is an difficult task in feature selection (FS) [1]. Fewer data, especially with particular fewer feature leads a better (e.g., more accurate) models. In machine learning (ML), data mining and statistics, evolutionary calculations (ECs) can solve optimization problems with some aspects of biological evolution. Evolutionary algorithms (EA) used the strategies of EC's. As EA follows EC's strategies thus like EC, EA also inspired by biological evolution. Some of its strategies include principal component (PC), particle swarm (PS), ant colony (AC), and genetic algorithm (GA) optimizations techniques [2]. The aim of this work is to apply two different ML methods (artificial neural networks and support vector machines), trained and structurally optimized by GA, and to compare the results with regression-based methods (logistic regression).

II. Literature Review

There are numerous works for GA to classified models across different domains. For example, Barrios et al. [3] trained the network using GA and later develops a classification technique for classification of breast cancer. In a research work, support vector machines (SVM) was use GA as well as to select a subset of input features. In the next step, the SVM classification technique was deployed to classified the defective or normal [4]. Beiko and Charlebois [5] applied GA to identify the best output by artificial neural networks (ANN) for classifications of DNA sequence. Karzynski et al. [6] applied GA to reduce data for ANN weight classification in the classifications of microarray. For example, Cho et al. [7] applied the ANN output by using GA in the predication and classification of cDNA microarray.

III. Methods

A. Study Design

A cross-sectional survey (Multiple indicator cluster surveys, 2012) data were used in this study, the response variable is the school attrition (Yes Vs No). Explanatory variables are included by the previously published research [8], [9].

B. Data Preparation

The whole data was equally separated into two subsets: training and validation (testing) data. The main purpose of the validation data is to identify problems in the classification models such as over fitting effects or premature termination of the training process. This study used standardization of Z-scores to removes potential outliers of the data matrix.

IV. Feature Selection Methods

FS is an important not only for the large data set but also because of dealings with multiple issues. Feature interactions occur frequently in many cases. Features can include many general way, sometimes, complex multi way interactions [10]. An appropriate feature can do multiple work simultaneously when used in parallel. However, FS is used to picked the key terms (features) [11], to create graphical insides in image analysis [12], and also to determine all important business indicators [13].

A. Genetic algorithm

The idea of GA was perceived by [14], as a technique to explore the variety of optimization complications [15]. Design of this method was inspired by observing the natural Darwinian evolutionary process and surviving the optimal principle [16]. Several parameters need to be implemented in GA (e.g. population size, mutation and crossover probability, and their interrelationships) [17]. To understand the theory behind genetic algorithms, one must first understand how recombination occurs. Reconnection requires two parents. Consider 1101001100101101 a string and other binary string yxyyxyxxyyyxyxxy. Where x and y are used to present 0 and 1. Recombination occurs using two "break-points" are as follows:

11010	01100101	101
yxyyx	yxxyyyxy	xxy

Swapping the fragments between the two parents produces the offspring: 11010yxxyyyxy101 and yxyyx01100101xxy.

It can also be used for training in logistic regression (LR), ANN, and SVM or determination of optimal performing structure. Since it is not using the error function of gradient to reach the best solution, it is also not sensitive to the local minima problem [18]. The step of GA is shown in Figure 1. First, randomly created coded was applied and divided into preliminary populations. Second, by calculating fitness function each was given fitness value. Additionally, perform a cross-over, which is an exchange of features from the selected subset, then introduce mutations, which are applied

randomly to the randomly selected features. Lastly, return to the second point. This continues until the exit criteria are found.

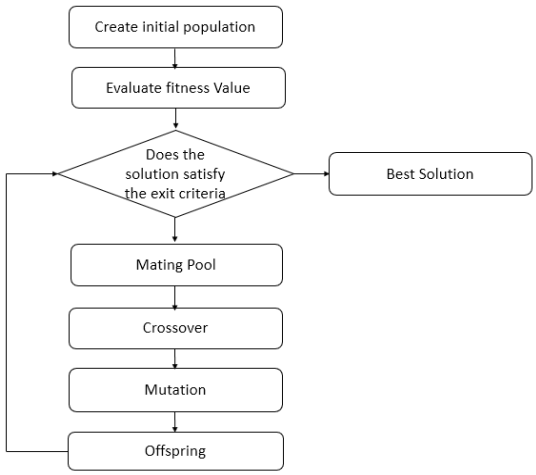

Figure. 1. Steps of genetic algorithm

Initial Population: The initial population size are the major features in GA. Initialization is known as the size of population which usually selected by the criteria of the problem by initially generated random trial.

Fitness computation: Fitness computation are also a key element of the classification. The fitness is well-defined and properly labelled returned by the input samples in our models.

Crossover: An encoded bit string and the exact number of bits before and after the point position is broken down and exchanged between parental variables.

Mutation: Mutations can be achieved with a small probability, for example, 0.001. A 100% mutation is that the whole dataset will be reformed, but a 0% means that no part of the dataset will be distorted.

B. Application of Genetic Algorithm

1) Initialization of population: Generate an initial population that is a randomly generated bit string of binary values.
2) Decoding: Decode (bit string) to find out which input variable to select.
3) Classification model: In this stage, applied classification models.
4) Fitness evaluation: Take the prediction accuracy of each model as the fitness for GA.
5) Stopping criterion: The loop should continue or exit was determined here.
6) Selection: Select the model to cross over using randomly selected variables of the population.
7) Crossover: A linear combination of two selection was apply by a crossover operator.
8) Mutation: Attach a new gene with the same operator and create a random slot number of cross-over variables 9 Replacement: Replace old variables with the two-best combination which generate from mutation for next step.
9) Loop: Go to Step 2.

C. Prediction Models

1) Artificial Neural Network: ANN provides an alternative method of interpreting and recognizing complex patterns in data sets. ANNs should be considered a form of converges that iterative itself to solve many classification problems.
2) Support Vector Machine: SVM is a method that can deal with noisy pattens in a large dataset. This approach is to find a function in a multidimensional space that can separate training data with known class labels.
3) Logistics Regression: LR is used for analysing binary response data. The outcome variable Y is denoted by 1 ("success") or 0 ("failure"). The logistic model of any variable, Y, is defined as,

$$1/(e^{(-Y)}+1) \quad (1)$$

D. Fitness Evaluation

The accuracy of the fitness values in GA:

$$\text{Accuracy} = (TP + TN)/ (TP + TN + FP + FN)$$

Where TP is true positive, FP is false positive, TN is true negative, and FN is false negative. True Positive (TP) are those items that are properly selected. False positives (FP) are those items that have been selected incorrectly. True Negative (TN) are items that should not be selected but are selected. False Negatives (FN) are items that should be selected but not selected.

$$\text{Recall}(R) = TP / (TP+FN)$$
$$\text{Precision } (P) = TP / (TP+FP)$$
$$\text{F1-score} = (2PR) / (P+R)$$

V. Results and Discussion

Performance criteria in this study was quantified by recall, precision and F1-score. The accuracy of how many selected items are acceptable and how many acceptable items are selected is called recall. A balanced blending of recall and precision is the F1-score. Performance metrics are also calculated in this study.

In this study, the results were obtained based on the GA configuration, given in Table 1. Carefully selected fitness functions enable GA to reduce classification errors from classification models. As a proof, the best fitness and mean fitness should be close to the standard because GA has reached the condition of termination. Stall Generation is the number of generations produced by GA after the latest upgrade of fitness value. In this study, GA was used single point crossover, mutation threshold (0.55), maximum iteration (10), population number (40) and number of offspring in each iteration (40). T. Liu et al. worked previously with the population size of 20 and after 200 iterations all algorithms are stopped [19]. The parameters used by the GA for all models are shown in Table I.

TABLE I. Parameters Used in GA

GA Parameter	Value
Crossover	Single point crossover
Mutation threshold	0.55
Maximum iteration	10
Number of population (Initial)	40

GA Parameter	Value
Number of offspring in each iteration	40

Using the GA feature selector in ANN classification techniques, 6 features were reported with a mean of 0.67. Interestingly, GA has selected 7 features in both SVM and LR models. Ahmed et al. found the same feature with the use of similar parameters [20]. Since the parameters in the GA configuration table can still be fine-tuned for better results, the GA method has a higher level of controllability, with the mean feature value being higher in all models Table II.

TABLE II. SELECTED FETURE

Feature	ANN	SVM	LR
Girls age (Year)	1	1	1
Marital status	1	1	1
Area	1	0	0
Divisions	1	1	1
Household wealth index	1	0	0
Religion	0	1	1
Household education	0	1	1
Mother alive	1	1	1
Father alive	0	1	1
Mean	0.67	0.78	0.78

The training parameters of ANN are shown in Table 3. After much experimental effort for an optimal model, we found that the number of hidden layers is 4 (6,9,9 and 4). The learning rate is 0.001. Regarding the learning rate chosen 0.01, the chosen 10 nodes for hidden layer and using 20% dataset for testing, have been reported in our previous study [21]. The selection of each parameter is shown in Table 3.

TABLE III. PARAMETERS USED IN ANN

ANN Parameter	Value
Hidden layer	4 [6,9,9,6]
Solver	adam
alpha	1e-5
activation	relu
output activation	softmax
Learning rate initialize	0.001
Maximum initialize	500

It is clear that, without GA the prediction accuracy is lowest in all method. Without GA the accuracy was 76.00%, 75.25% and 74.92% in ANN, SVM and LR, respectively. On the other hand, With GA feature section method, the accuracy was 76.50%, 76.00% and 75.83% in ANN, SVM and LR, respectively (Table 4). Kim, used similar findings with us in their study. It is clear that, all method shows highest accuracy with GA [22].

TABLE IV. CLASSSIFICATION ACCURACY WITH AND WITHOUT GA

Classification Method	Without GA	With GA	Training time
ANN	76.00%	76.50%	2.864
SVM	75.25%	76.00%	0.603
LR	74.92%	75.83%	0.043

We got the best recall of 95.39% in LR model which is better for this model than 50% and F1 score is highest on LR model and all method showed results above 50%. That means, all model fit best to our data.

TABLE V. RESULTS OF DROPOUT CLASSIFICATION USING GA ON DIFFERENT METHODS

Methods	Recall (R) (%)	Precession (P) (%)	F1-score (%)
ANN	92.35	77.65	84.36
SVM	93.67	76.59	84.27
LR	95.39	75.72	84.42

We got the best recall of 95.39% in LR model which is better for this model than 50% and F1 score is highest on LR model and all method showed results above 50%. That means, all model fit best to our data.

VI. CONCLUSION

In this study, we developed an individual classification method and GA model for classifying school attrition data. The results of the study show that the genetic algorithm got a good result for a small range of initial parameters. In most cases, the difference between the accuracy of the classification reported with and without GA is very small. Overall, the GA feature selectors created better classification accuracy without imply to this method. The main advantage of this approach is that it belongs to the field of controllable because GA can be secured for better results all the time by changing the fitness functions.

REFERENCES

[1] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016, doi: 10.1109/TEVC.2015.2504420.

[2] B. de la Iglesia, "Evolutionary computation for feature selection in classification problems," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 6, pp. 381–407, Nov. 2013, doi: 10.1002/widm.1106.

[3] D. Barrios, A. Carrascal, D. Manrique, and J. Ríos, "Cooperative binary-real coded genetic algorithms for generating and adapting artificial neural networks," *Neural Comput. Appl.*, vol. 12, no. 2, pp. 49–60, Nov. 2003, doi: 10.1007/s00521-003-0364-1.

[4] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, "Bearing Fault Detection Using Artificial Neural Networks and Genetic Algorithm," *EURASIP J. Appl. Signal Processing*, vol. 2004, no. 3, pp. 366–377, Mar. 2004, doi: 10.1155/S1110865704310085.

[5] R. G. Beiko and R. L. Charlebois, “GANN: Genetic algorithm neural networks for the detection of conserved combinations of features in DNA,” *BMC Bioinformatics*, vol. 6, no. 1, p. 36, Feb. 2005, doi: 10.1186/1471-2105-6-36.

[6] M. Karzynski, Á. Mateos, J. Herrero, and J. Dopazo, “Using a genetic algorithm and a perceptron for feature selection and supervised class learning in DNA microarray data,” *Artif. Intell. Rev.*, vol. 20, no. 1–2, pp. 39–51, Oct. 2003, doi: 10.1023/A:1026032530166.

[7] H. S. Cho, T. S. Kim, J. W. Wee, S. M. Jeon, and C. H. Lee, “cDNA microarray data based classification of cancers using neural networks and genetic algorithms,” in *2003 Nanotechnology Conference and Trade Show - Nanotech 2003*, 2003, vol. 1, pp. 28–31, Accessed: Aug. 07, 2020. [Online]. Available: https://briefs.techconnect.org/papers/cdna-microarray-data-based-classification-of-cancers-using-neural-networks-and-genetic-algorithms/.

[8] M. N. Hasan, “A Comparison of Logistic Regression and Linear Discriminant Analysis in Predicting of Female Students Attrition from School in Bangladesh,” Dec. 2019, doi: 10.1109/EICT48899.2019.9068776.

[9] M. N. Hasan, “Factors Associated with Attrition of Girls Students from School in Bangladesh,” *J. Sci. Res.*, vol. 12, no. 1, pp. 29–38, Jan. 2020, doi: 10.3329/jsr.v12i1.41579.

[10] I. Guyon and A. M. De, “An Introduction to Variable and Feature Selection André Elisseeff,” 2003.

[11] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, “Text feature selection using ant colony optimization,” *Expert Syst. Appl.*, vol. 36, no. 3 PART 2, pp. 6843–6853, Apr. 2009, doi: 10.1016/j.eswa.2008.08.022.

[12] A. Ghosh, A. Datta, and S. Ghosh, “Self-adaptive differential evolution for feature selection in hyperspectral image data,” *Appl. Soft Comput. J.*, vol. 13, no. 4, pp. 1969–1977, Apr. 2013, doi: 10.1016/j.asoc.2012.11.042.

[13] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005, doi: 10.1109/TKDE.2005.66.

[14] J. H. Holland, “Adaptation in Natural and Artificial Systems | The MIT Press,” *The University of Michigan Press*, 1975. https://mitpress.mit.edu/books/adaptation-natural-and-artificial-systems (accessed Jul. 31, 2020).

[15] H. Chiroma *et al.*, “Neural networks optimization through genetic algorithm searches: A review,” *Applied Mathematics and Information Sciences*, vol. 11, no. 6. pp. 1543–1564, 2017, doi: 10.18576/amis/110602.

[16] M. Hamdan, “A heterogeneous framework for the global parallelisation of genetic algorithms,” *Int. Arab J. Inf. Technol.*, vol. 5, no. 2, pp. 192–199, 2008, Accessed: Jul. 31, 2020. [Online]. Available: https://www.researchgate.net/publication/220413569_A_Heterogeneous_Framework_for_the_Global_Parallelisation_of_Genetic_Algorithms.

[17] C. T. Capraro, I. Bradaric, G. T. Capraro, and T. K. Lue, “Using genetic algorithms for radar waveform selection,” 2008, doi: 10.1109/RADAR.2008.4720947.

[18] “BP Neural Network Algorithm Optimized by Genetic Algorithm and Its Simulation,” *Int. J. Comput. Sci. Issues*, vol. 10, no. 1, pp. 516–519, 2013, Accessed: Jul. 31, 2020. [Online]. Available: https://www.researchgate.net/publication/303102486_BP_Neural_Network_Algorithm_Optimized_by_Genetic_Algorithm_and_Its_Simulation.

[19] T. Liu, H. Zhang, H. Zhang, and A. Zhou, “Information Fusion in Offspring Generation: A Case Study in Gene Expression Programming,” *IEEE Access*, vol. 8, pp. 74782–74792, 2020, doi: 10.1109/ACCESS.2020.2988587.

[20] F. Ahmad, N. A. Mat-Isa, Z. Hussain, R. Boudville, and M. K. Osman, “Genetic Algorithm - Artificial Neural Network (GA-ANN) hybrid intelligence for cancer diagnosis,” in *Proceedings - 2nd International Conference on Computational Intelligence, Communication Systems and Networks, CICSyN 2010*, 2010, pp. 78–83, doi: 10.1109/CICSyN.2010.46.

[21] Y. T. Chang, J. Lin, J. S. Shieh, and M. F. Abbod, “Optimization the initial weights of artificial neural networks via genetic algorithm applied to hip bone fracture prediction,” *Adv. Fuzzy Syst.*, 2012, doi: 10.1155/2012/951247.

[22] K. Kim, “Artificial neural networks with evolutionary instance selection for financial forecasting,” *Expert Syst. Appl.*, vol. 30, no. 3, pp. 519–526, Apr. 2006, doi: 10.1016/j.eswa.2005.10.007.