Machine learning of tax avoidance detection based on hybrid metaheuristics algorithms

Suraya Masrom[1], Rahayu Abdul Rahman[2], Masurah Mohamad[1], Abdullah Sani Abd Rahman[3], Norhayati Baharun[1]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Malaysia
[2]Faculty of Accountancy, Universiti Teknologi MARA, Perak Branch, Malaysia
[3]Faculty of Sciences and Information Technology, Universiti Teknologi PETRONAS, Perak, Malaysia

Article Info

Article history:

Received Aug 9, 2021
Revised May 31, 2022
Accepted Jun 18, 2022

Keywords:

Automated features selection
Genetic algorithm
Machine learning
Particle swarm optimization
Tax avoidance

ABSTRACT

This paper addresses the performances of machine learning classification models for the detection of tax avoidance problems. The machine learning models employed automated features selection with hybrid two metaheuristics algorithms namely particle swarm optimization (PSO) and genetic algorithm (GA). Dealing with a real dataset on the tax avoidance cases among companies in Malaysia, has created a stumbling block for the conventional machine learning models to achieve higher accuracy in the detection process as the associations among all of the features in the datasets are extremely low. This paper presents a hybrid meta-heuristic between PSO and adaptive GA operators for the optimization of features selection in the machine learning models. The hybrid PSO-GA has been designed to employ three adaptive GA operators hence three groups of features selection will be generated. The three groups of features selection were used in random forest (RF), k-nearest neighbor (k-NN), and support vector machine (SVM). The results showed that most models that used PSO-GA hybrids have achieved better accuracy than the conventional approach (using all features from the dataset). The most accurate machine learning model was SVM, which used a PSO-GA hybrid with adaptive GA mutation.

Corresponding Author:

Rahayu Abdul Rahman
Faculty of Accountancy, Universiti Teknologi MARA
Perak Branch, Malaysia
Email: rahay916@uitm.edu.my

1. INTRODUCTION

Particle swarm optimization (PSO) [1], [2] and genetic algorithm (GA) [3], [4] are two common metaheuristics algorithms that have been widely used to solve different applications of optimization problems. To date, the hybridization between PSO and GA has been very useful to enable high-performance searching of the optimization space. Recently, with the broad utilization of machine learning [5]–[7], PSO and GA have given their important roles to help the models in achieving highly accurate results for classification, regression, clustering, and forecasting models [8]. PSO and GA can be used in many ways for the machine learning models, either individually [9] or in combination [10] to automate the important tasks in the machine learning pipelines. This paper presents the roles of PSO-GA hybridization as an optimization tool for machine learning features selection.

It is anticipated in this research that the PSO-GA hybridization [11], [12] can be very useful to be deployed for the application of machine learning, mainly for the problem that involved real cases dataset. The results of machine learning models from our previous studies in [13], [14] on the tax avoidance detection were undesirable due to the problem of very weak correlation among the dataset features. Although data engineering [15] exercises can be conducted to improve the knowledge extrapolations from the dataset it is expected that automating the features selection without original data manipulation can be more helpful. The contributions of this paper are multifaceted. Firstly, the machine learning models can be used to resolve the problem of tax avoidance issue that occurs in Malaysia. Although corporate tax is the highest contributor to government revenues, it can signify the biggest burden of the cost incurred by the firms. Thus, managers attempt to minimize the tax liability by using various legal and illegal strategies including tax avoidance plans. Therefore, the development of the corporate tax avoidance model has long been seen as significant to the tax authorities and business community. Research in machine learning classification models for detecting tax avoidance is infrequently in the current literature. This work was initially inspired by the machine learning tax avoidance prediction research conducted by [16] that used logistic regression, decision tree, and random forest machine learning algorithms. Based on our previous studies, using automated machine learning has given more advantages compared to manual machine learning configurations [14] but this approach only used genetic programming (GP) algorithm to optimize the features selections of the tax avoidance dataset. To get a research report on PSO with adaptive GA operators is difficult from the literature neither in tax avoidance nor in other machine learning applications.

The second contribution of this paper is on the deployment of hybridization approach in machine learning. As features selection is critical in machine learning, different approaches of PSO-GA features selection have been intensively tested in different machine learning models. Additionally, the problem that exists in the collected dataset related to the tax avoidance cases among government-link companies (GLCs) in Malaysia is very low correlations among the independent variables (IVs). Removal of some features from the IVs without adequate research might lower the classification accuracy of the machine learning models. Therefore, the optimization of features selection based on the proposed PSO-GA hybrids was expected to be useful in the tax avoidance application.

Thirdly, this paper provides a research report that extends the study on a problem related to PSO premature convergence. The results reported in this paper present the benefits of hybridizing adaptive GA operators in PSO to resolve the premature convergence in achieving the most optimal results. The single PSO faced an imbalanced problem between the exploration and exploitation searching direction [17]. The GA has some operators when used appropriately in the PSO, and can be useful to enable exploration and exploitation steadiness.

2. RESEARCH METHOD

2.1. PSO, GA, and PSO-GA hybrids

PSO and GA are among the popular nature-inspired metaheuristics algorithms that use the current state of the search performances to determine the next search direction. Both algorithms are from the population-based meta-heuristics [18], [19]. The PSO mimics the cognitive and social behaviors of birds flocking [20] while GA simulates the evolution of creatures [21].

In PSO, the particles consists of a D-dimensional position vector and a D-dimensional velocity vector as shown in Figure 1. Each position of an *i*th particle can be represented as x while the *i*th particle velocity can be denoted as *v*. In each iteration *t*, every velocity of each particle in the population has to move towards the best fitness based on the previous experience. The calculation to update the next velocity of each particle *i* is determined by (1).

$$v_{it} = v_{i(t-1)} + (c_1 * r_1 * (pbest - x_{i(t-1)})) + (c_2 * r_2 * (gbest - x_{i(t-1)}) \tag{1}$$

where c_1 and c_2 are the acceleration coefficient with positive constants and the r_1 and r_2 are two different numbers generated randomly between 0 and 1.

The (1) affects accelerating particles toward the weighted sum of its personal best position *pbest* and global best position of the swarm (*gbest*) from the previous particle position $x_{i(t-1)}$. After updating each particle velocity, (2) will be implemented for updating the position of each particle *i* at each iteration *t*.

$$x_{it} = x_{it} + v_{it} \tag{2}$$

where x_{it} is the updated position at the current iteration and v_{it} is the new velocity that was calculated from (1). The proses of updating each particle's velocity and position will be repeated until the optimization objective has been met or until the maximum iteration has been reached. Early or premature convergence in PSO is a common problem that occurred when the PSO ends this process before reaching the optimal objective. To resolve the problem of premature convergence, hybridization with GA operators (selection, crossover, and mutation) is expected to be helpful in PSO.

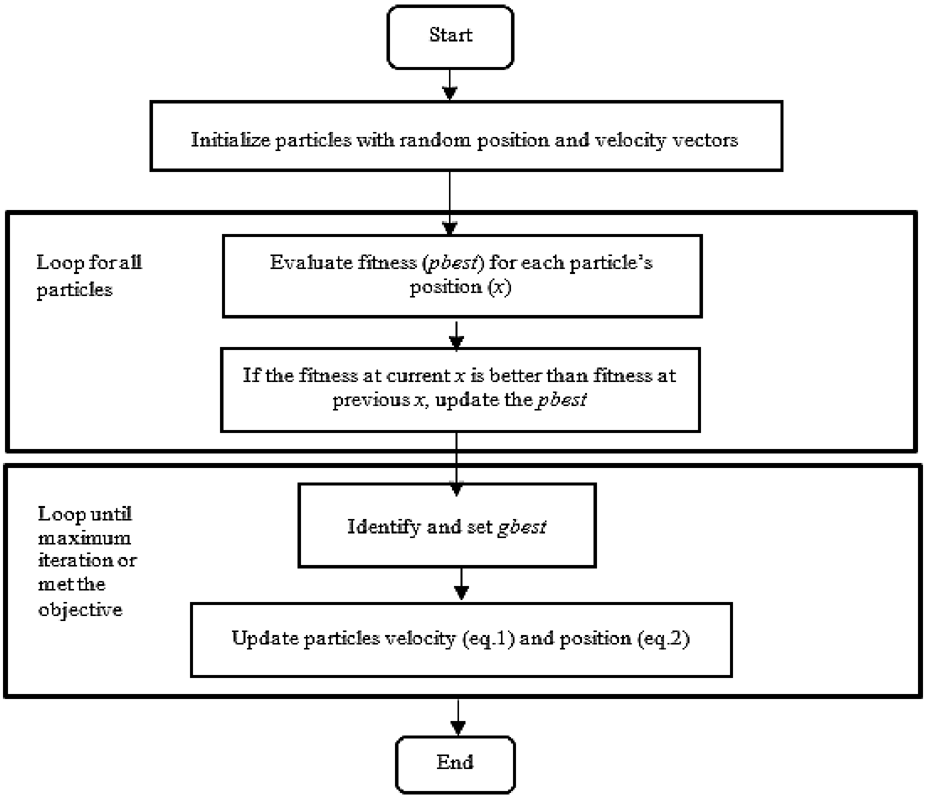

Start
Initialize particles with random position and velocity vectors
Loop for all particles
Evaluate fitness (*pbest*) for each particle's position (*x*)
If the fitness at current *x* is better than fitness at previous *x*, update the *pbest*
Loop until maximum iteration or met the objective
Identify and set *gbest*
Update particles velocity (eq.1) and position (eq.2)
End

Figure 1. The flowchart of the PSO algorithm

Different from PSO which implements individual repositioning, GA uses individual reproduction. As depicted in Figure 2, a proportion of the existing population in GA will be chosen to produce a new generation during each successive reproduction. The proportion of individual of chromosome from the current population is implemented with a fitness-based selection.

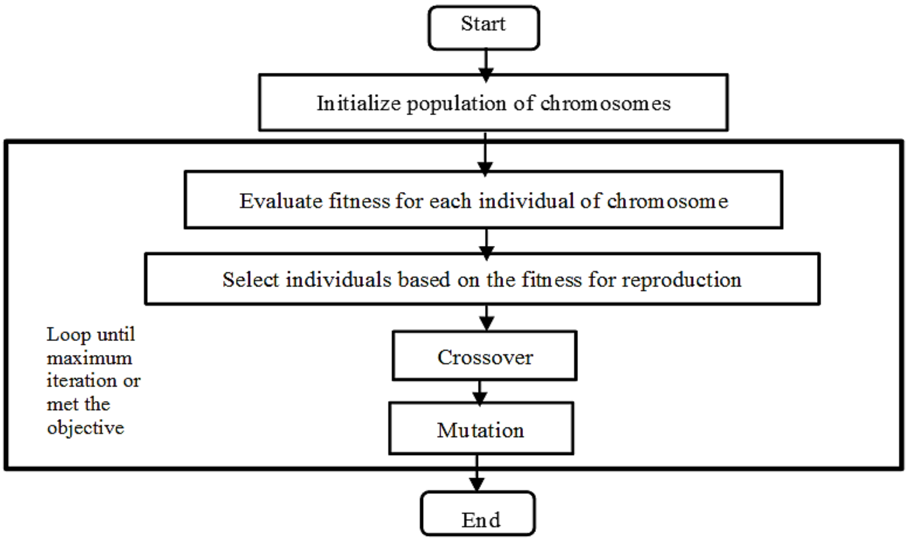
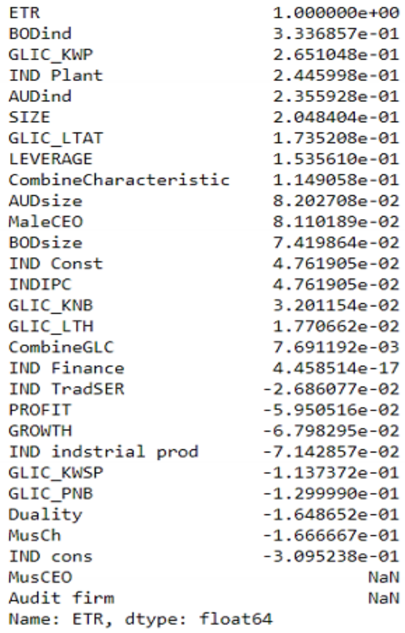

Start
Initialize population of chromosomes
Evaluate fitness for each individual of chromosome
Select individuals based on the fitness for reproduction
Loop until maximum iteration or met the objective
Crossover
Mutation
End

Figure 2. The flowchart of the GA algorithm

The crossover is an operation that combines two chromosomes/parents to produce a new chromosome/offspring. The new offspring should be better than the parents. Moreover, the mutation is the GA operator that alters one or more gene values in a chromosome from its current state. Hybridization of PSO with GA is not a new technique in meta-heuristics algorithms. They can be combined in a variety of approaches either with low-level or high-level hybridizations [22]. High-level hybridization allows them to be linked without internal algorithm modification such as executing the algorithm separately in parallel or sequences. Otherwise, the low-level hybridization involves internal flow amendment of the algorithms individually or in combinations. The proposed PSO-GA in this paper is a low-level hybridization.

2.2. The tax avoidance dataset

The dataset was collected from the GLCs in Malaysia from the period between 2010 to 2016. This study used effective tax rates (ETR) to denote the tax avoidance, which was formulated based on the ratio of the total tax expenses with the total income before tax. The GLCs companies can be classified as tax avoidance firms if the ETR is smaller than the corporate statutory tax rates [13]. A detailed description of the features can be referred in [13]. As depicted in Figure 3, all the features have a very weak correlation to the ETR mainly from the finance indicator (IND Finance). Particularly, the Muslim Chief Executive Officer (MusCEO) and Audit Firm variables have no significance at all to the ETR. However, in machine learning models, all the features even with low or zero correlation to the ETR, when combined may contribute some degree of knowledge to the classification models. But the question raised is related to which feature combination has optimal contribution to help the models produce the best accuracy results. Manual features selection is quite impractical to involve all the possible sets of combinations. Therefore, automated features selection was implemented for the machine learning models.



```
ETR                           1.000000e+00
BODind                        3.336857e-01
GLIC_KWP                      2.651048e-01
IND Plant                     2.445998e-01
AUDind                        2.355928e-01
SIZE                          2.048404e-01
GLIC_LTAT                     1.735208e-01
LEVERAGE                      1.535610e-01
CombineCharacteristic         1.149058e-01
AUDsize                       8.202708e-02
MaleCEO                       8.110189e-02
BODsize                       7.419864e-02
IND Const                     4.761905e-02
INDIPC                        4.761905e-02
GLIC_KNB                      3.201154e-02
GLIC_LTH                      1.770662e-02
CombineGLC                    7.691192e-03
IND Finance                   4.458514e-17
IND TradSER                  -2.686077e-02
PROFIT                       -5.950516e-02
GROWTH                       -6.798295e-02
IND indstrial prod           -7.142857e-02
GLIC_KWSP                    -1.137372e-01
GLIC_PNB                     -1.299990e-01
Duality                      -1.648652e-01
MusCh                        -1.666667e-01
IND cons                     -3.095238e-01
MusCEO                                 NaN
Audit firm                             NaN
Name: ETR, dtype: float64
```

Figure 3. The pearson correlation of each feature to the ETR

2.3. The proposed PSO-GA hybrids

After the general best fitness of the PSO has been identified in a particular loop, the selection operator from GA will be executed. Then, the adaptive crossover, adaptive mutation, or both adaptive crossover mutation will be implemented. Figure 4 presents the flowchart of the proposed PSO-GA hybrids that used adaptive GA operators.

To examine the effect of each adaptive operator, three types of PSO-GA hybrids for the features selection have been developed. The first approach used both adaptive operators (PSO-ACM), the second used adaptive crossover (PSO-AC), and the third used adaptive mutation (PSO-AM). Based on Figure 5, after randomly selecting two particles, the adaptive crossover will be implemented followed by adaptive mutation for the PSO-ACM.

In the PSO-AC, two particles will be selected to generate new off-string particles while in the PSO-AM, only one particle is selected for adaptive mutation. Figure 5 presents the pseudocodes of PSO-AC. The adaptive crossover in PSO-AC executes the crossover of two particles that are randomly selected according to the adaptive crossover probability $C_{p.}$ The C_p of all particles in line 2 used the adaptive technique introduced by in Qin *et al.* [23], which introduced the formulation of individual search ability (ISA). As given in (3), the ISA formula calculates the ratio of distance for particle *i* in the *d* dimension.

$$ISA_{i(d)} = \frac{|x_{i(d)-} \; pbest_{i(t)}|}{|pbest_{i(t)} - gbest_t \,| + \varepsilon} \tag{3}$$

where $x_{ia\,(d)}$ denotes the recent position of the *i*th particle. The personal best position of the *i*th particle in the current iteration is denoted as $pbest_{i(t)}$. Then, $gbest_t$ is the latest global best position of the entire swarm, a nd ε is a positive constant that is nearest to zero. The numerator is a distance from the personal best to the current position while denominator is the result of distance from the global best to the personal best.

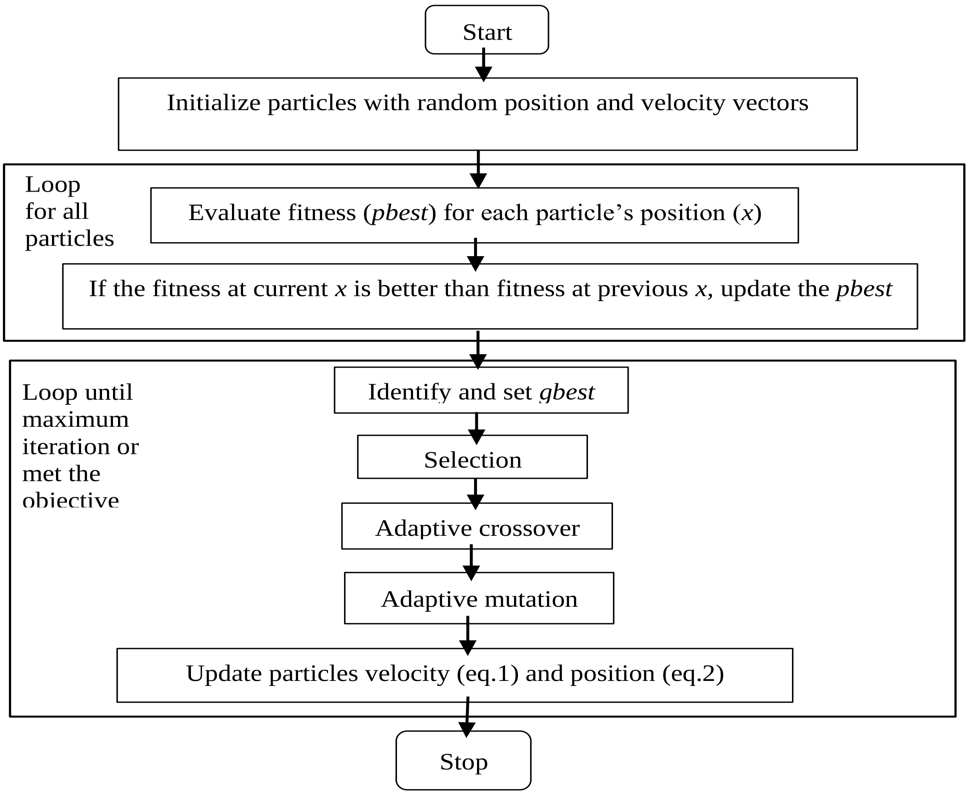
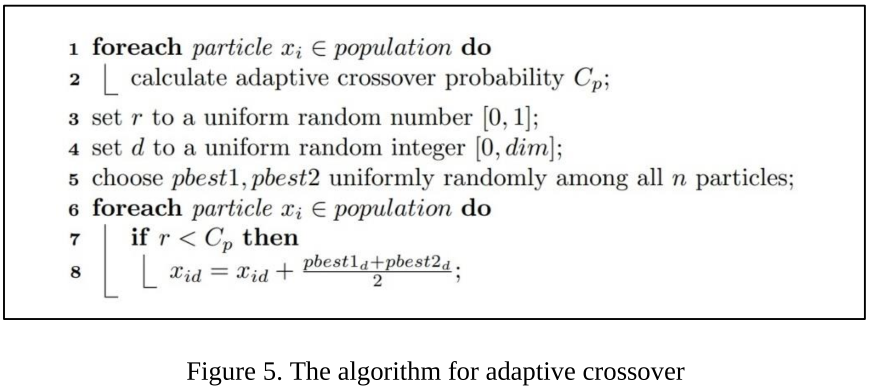

Start
Initialize particles with random position and velocity vectors
Loop for all particles
Evaluate fitness (*pbest*) for each particle's position (*x*)
If the fitness at current *x* is better than fitness at previous *x*, update the *pbest*
Loop until maximum iteration or met the objective
Identify and set *gbest*
Selection
Adaptive crossover
Adaptive mutation
Update particles velocity (eq.1) and position (eq.2)
Stop

Figure 4. The proposed PSO-GA hybrids



```
1 foreach particle $x_i \in population$ do
2     calculate adaptive crossover probability $C_p$;
3 set $r$ to a uniform random number $[0, 1]$;
4 set $d$ to a uniform random integer $[0, dim]$;
5 choose $pbest1, pbest2$ uniformly randomly among all $n$ particles;
6 foreach particle $x_i \in population$ do
7     if $r < C_p$ then
8         $x_{id} = x_{id} + \frac{pbest1_d + pbest2_d}{2}$;
```

Figure 5. The algorithm for adaptive crossover

The threshold *r* value has been set with a random number within an interval of zero and one, which the value has to be compared with the particle's probability C_p of crossover in deciding whether this particle at its random position *d* should be modified using the crossover operator or not. The crossover operator in line 8 was adopted from [24]. Furthermore, the adaptive mutation also used the ISA scheme for the mutation probability M_p to decide which particles should be mutated. As shown in Figure 4 at line 5, the mutation only chooses one *pbest* particle from the uniformly random *n* particles. Furthermore, the operation in line 8 in Figure 5 was replaced with (4).

$$x_{i(d)} = x_{i(d)} + Gaussian(\sigma) \tag{4}$$

where the Gaussian function [25] returns a random value from the range of the particle dimension and the σ value is in between 0.1 times of the particle dimension. Referring back to Figure 4, the population of the PSO is a set of particles that represent the Malaysian GLCs records listed in the year 2010-2016. Figure 6 is the particle representation for the features.

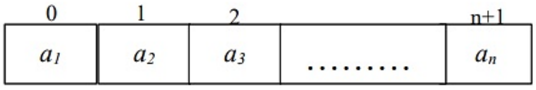


Figure 6. The PSO solution representation

The dimension size of each particle is the total number of features (28). Each particle stores the features' column id ($a_1..a_n$) of the dataset (converted to a data frame in Python). The objective fitness function is to get the maximum accuracy value from the machine learning that will use the randomized features selection. When the fitness is better than the previous *pbest*, the latest position vector is saved for the particle. If the fitness is better than the global best fitness, then the position vector *x* is also saved for the global best *gbest*. Finally, the particle's velocity and position are updated with (1) and (2) until the termination condition is satisfied. The general experiment setting is given in Table 1.

Table 1. General experiment setting for the PSO hybrid with adaptive GA

Attribute	Value
Particles number, *n*	10, 20, 30
Particle dimension, *dim*	Number of features (28)
Personal learning rate, c_1	0.9
Social learning rate, c_2	0.9
Iterations number, *i*	100-1000

Regardless of each algorithm, the number of function evaluations is the number of particles multiply the number of iterations. Therefore, for the 30 particles with 1,000 iterations, the total number of evaluations is 30,000. It is important to identify the suitable number of particles and iterations based on the input dataset. The preliminary experiments have been conducted with particles 10, 20 and 30. The iterations number to be observed is from 100 to 1000. Figure 7 presents the implementation of features selection optimization with the proposed PSO-GA in the machine learning classification model for tax avoidance detection.

The machine learning algorithms used in this research were support vector machine (SVM), random forest (RF), and K-nearest neighbor (k-NN). Table 2 lists the classification machine learning algorithms with the parameters setting implemented in Python codes. The hardware for running the machine learning models is a Lenovo notebook Intel i7 7th generation processor with 16 GB RAM.

Table 2. The machine learning algorithms and their parameters configuration

Machine learning call and parameters parsing in Python
svc = SVC(C=2,kernel='poly')
knn = KNeighborsClassifier (n_neighbors = 8)
rf= RandomForestClassifier(estimator's=100)

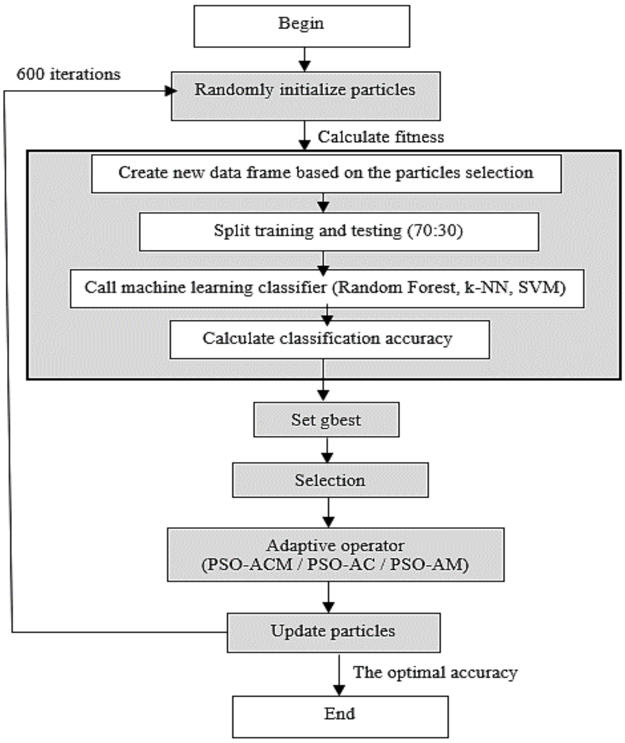

Begin
600 iterations
Randomly initialize particles
Calculate fitness
Create new data frame based on the particles selection
Split training and testing (70:30)
Call machine learning classifier (Random Forest, k-NN, SVM)
Calculate classification accuracy
Set gbest
Selection
Adaptive operator
(PSO-ACM / PSO-AC / PSO-AM)
Update particles
The optimal accuracy
End

Figure 7. The PSO-GA features selection in the machine learning models

3. RESULTS AND DISCUSSION

The results are divided into two. The first results are in plotted graphs to present the accuracy of each features selection approach (PSO-AM, PSO-AC, PSO-ACM) in the machine learning models according to the number of particles (population size at *n=10,n=20, n=30*). The *y* axis is the accuracy percentage of models and the *x-axis* presents several iterations (100-1000). From the graphs, the suitable number of particles *n* that can generate the most accurate result and the number of iterations *i* of when the algorithm was stagnated can be depicted. Furthermore, the second results compare the accuracy results of machine learning models that used the three features selection approaches at the selected *n* and *i*, the single PSO without hybridization, and the conventional manual features selection (all features). Figure 8 presents the graphs of accuracy results from the PSO-AM features selection approach in the SVM, k-NN, and RF at a different number of iterations.

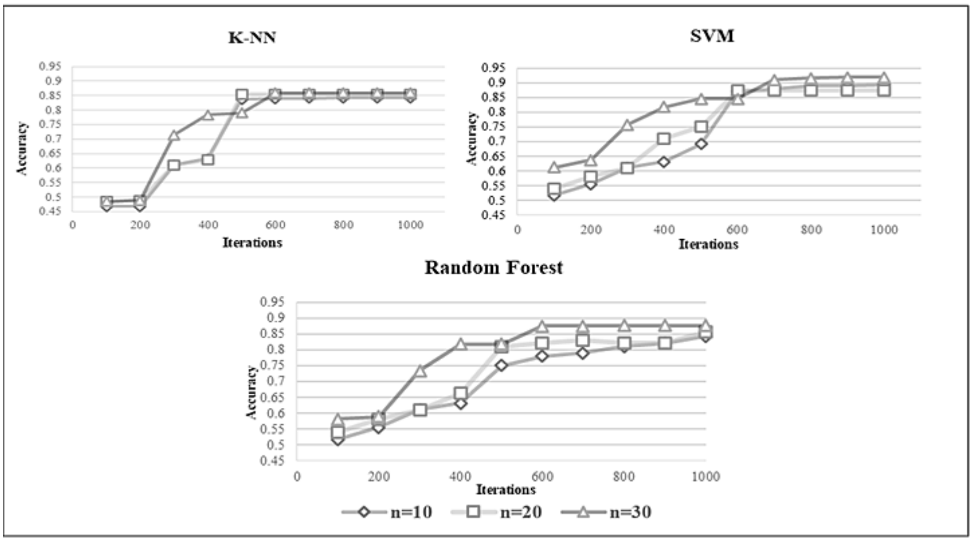


Figure 8. The PSO-AM features selection in the k-NN, SVM, and RF

The results in Figure 8 show that most of the machine learning models converged for optimal accuracy at 600 iterations. It can be seen in all the machine learning models that the convergences have occurred earlier with small populations (n=10, n=20) compared to n=30 mainly in SVM and k-NN. The 30 number of particles in k-NN and RF, have shown prior convergence at 600 iterations compared to 700 iterations in SVM. In RF, the algorithm tends to converge at 500 iterations with small populations but slight improvements have been shown at 1,000 iterations. Even with n=10, a good accuracy level (more than 75%) can be achieved by using the PSO-AM features selection mainly in k-NN and SVM. When the n=30, all the models have produced above 85% of accuracy level after 600 number iterations.

Furthermore, Figure 9 presents the convergence rates of PSO-AC. By using adaptive crossover in PSO-AC features selection, most of the machine learning models even with n=30 have shown faster convergences (less than 600) than the adaptive mutation (PSO-AM) but it decreased the accuracy result of all machine learning models. In RF, populations n=10 and n=20 have started to achieve the stagnation level at 300 iterations with an accuracy of less than 60%. Overall, all models with PSO-AC generated accuracy results of less than 80% when reached stagnation levels at iterations less than 600. In k-NN, even at the late convergence (iterations 600) has occurred for n=30, the optimal accuracy remained lower than the PSO-AM (less than 80%). Therefore, hybridizing the PSO with adaptive crossover does not appear to have much benefit in the accuracy of the machine learning models. However, as presented in Figure 10, the inclusion of adaptive mutation together with the adaptive crossover can be useful to improve the results of PSO-AC.

The convergence rates presented in Figure 10 are between 500 to 700 iterations (SVM and RF took 500 while k-NN with 700), and most of them were longer than PSO-AM but shorter than PSO-AC. Most of the models have generated better accuracy than PSO-AC (more than 85%). In SVM, n=20 and n=30 present the same number of iterations for the convergence at 500 and slightly late (600) with n=10. In k-NN, the models converged at 700 for all numbers of particles and the accuracy results are above 80%. In RF, the models have taken 500 iterations to converge with slight differences in accuracy levels (minimum above 75% and maximum below 90%).

In general, most of the machine learning models with the PSO-GA hybrids can achieve optimal accuracy results with 30 particles at iteration less or equal to 600. The results of each model with this configuration can be achieved within a reasonable time (less than 60 minutes). Therefore, it is important to compare the results of all models with the different PSO-GA hybrid approaches at this setting (n=30, i=600) with the manual approach that used all features. Additionally, it is interesting to compare all results with the model that used a single PSO without hybridization as listed in Table 3.

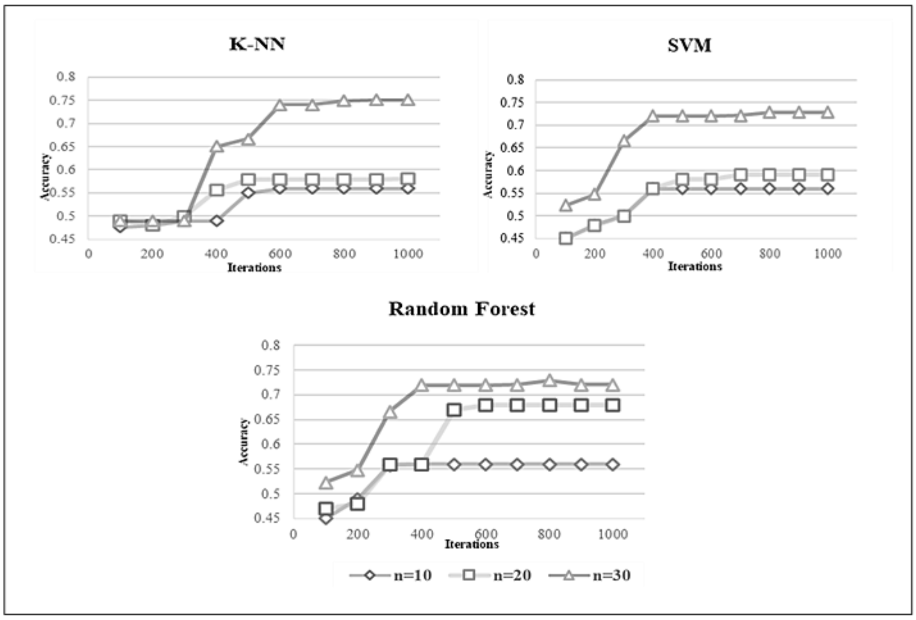


Figure 9. The PSO-AC features selection in the k-NN, SVM, and Random Forest

Generally, most models that used PSO-GA hybrids have achieved better accuracy than the conventional ones that used all features as well as with the use of a single PSO. SVM model can reach up to

91% of accuracy with the use of PSO-AM, which is the most outperformed machine learning model for the tax avoidance problem. Including both mutation and crossover PSO-ACM improved the performance of the machine learning models that used crossover (PSO-AC) in all cases. Moreover, the use of single PSO, as well as PSO-AC, appeared to be no benefit to the machine learning models as the accuracy results are lesser than the conventional approach that used all features.

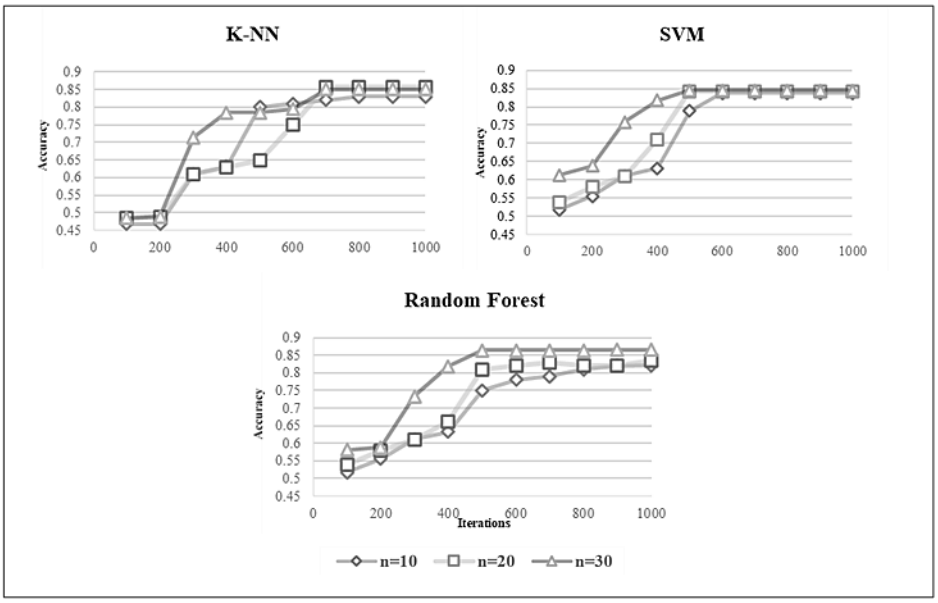


Figure 10. PSO-ACM features selection in the k-NN, SVM, and RF

Table 3. The accuracy results of the machine learning with different features selection approaches

	PSO-AM	PSO-AC	PSO-ACM	Single PSO	All features
SVM	0.91	0.72	0.85	0.75	0.83
k-NN	0.86	0.75	0.75	0.70	0.72
Random Forest	0.88	0.72	0.86	0.78	0.83

4. CONCLUSION

To the best of our knowledge on the state-of-the-art of features selection based on PSO-GA, the adaption of adaptive parameterization has not yet been introduced in the machine learning model. The literature revealed that adaptive GA operators can make significant improvements regarding the PSO algorithm’s performance when applied to many kinds of real-life problems. Therefore, this study attempts to introduce several PSO hybrids combined with adaptive GA operators. The results from the proposed approaches present additional advantages to the machine learning models when tested with a real dataset of tax avoidance problems. This research has opened up many research opportunities related to automated features selection of machine learning models as well as to the tax avoidance problem. For example, the proposed PSO-GA can be furtherly improved with different approaches to parameterizations within the PSO and the GA. Besides adaptive parameterizations, time-vary formulations can be other options for calculating the mutation and crossover rates. Furthermore, these varieties of PSO-GA hybrid approaches can be used not just for the features selections but also for the automated parameters tuning of the machine learning models. On the tax avoidance problem, different factors of ETR and different types of firms or businesses are also important to be explored. Last but not least, the variety of approaches from the PSO-GA hybrid can be applied or tested in any kind of application or problem domain.

ACKNOWLEDGEMENTS

We acknowledge the Ministry of Higher Education Malaysia and the Universiti Teknologi MARA that have given full support to this project under the FRGS grant number 600-IRMI/FRGS 5/3 (208/2019).

REFERENCES

[1] S. Sengupta, S. Basak, and R. A. Peters, "Particle Swarm Optimization: A survey of historical and recent developments with hybridization perspectives," *Mach. Learn. Knowl. Extr.*, vol. 1, no. 1, pp. 157–191, 2018.

[2] P. Matrenin *et al.*, "Generalized swarm intelligence algorithms with domain-specific heuristics," *IAES Int. J. Artif. Intell.*, vol. 10, no. 1, p. 157, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp157-165.

[3] K. Kamil, K. H. Chong, H. Hashim, and S. A. Shaaya, "A multiple mitosis genetic algorithm," *IAES Int. J. Artif. Intell.*, vol. 8, no. 3, pp. 252–258, Dec. 2019, doi: 10.11591/ijai.v8.i3.pp252-258.

[4] S. Mirjalili, J. Song Dong, A. S. Sadiq, and H. Faris, "Genetic algorithm: Theory, literature review, and application in image reconstruction," *Nature-inspired Optim.*, pp. 69–85, 2020.

[5] N. Razali, S. Ismail, and A. Mustapha, "Machine learning approach for flood risks prediction," *IAES Int. J. Artif. Intell.*, vol. 9, no. 1, pp. 73–80, Mar. 2020, doi: 10.11591/ijai.v9.i1.pp73-80.

[6] V. S. Padala, K. Gandhi, and P. Dasari, "Machine learning: the new language for applications," *IAES Int. J. Artif. Intell.*, vol. 8, no. 4, pp. 411–412, Dec. 2019, doi: 10.11591/ijai.v8.i4.pp411-421.

[7] A. M. Abdu, M. M. M. Mokji, and U. U. U. Sheikh, "Machine learning for plant disease detection: an investigative comparison between support vector machine and deep learning," *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, pp. 670–683, Dec. 2020, doi: 10.11591/ijai.v9.i4.pp670-683.

[8] O. Almomani, "A feature selection model for network intrusion detection system based on PSO, GWO, FFA and GA algorithms," *Symmetry (Basel).*, vol. 12, no. 6, p. 1046, Jun. 2020, doi: 10.3390/sym12061046.

[9] T. Khadhraoui, S. Ktata, F. Benzarti, and H. Amiri, "Features selection based on modified PSO algorithm for 2D face recognition," in *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*, Mar. 2016, pp. 99–104, doi: 10.1109/cgiv.2016.28.

[10] A. Benvidi, S. Abbasi, S. Gharaghani, M. D. Tezerjani, and S. Masoum, "Spectrophotometric determination of synthetic colorants using PSO--GA-ANN," *Food Chem.*, vol. 220, pp. 377–384, Apr. 2017, doi: 10.1016/j.foodchem.2016.10.010.

[11] A. M. Manasrah and H. B. Ali, "Workflow scheduling using hybrid GA-PSO algorithm in cloud computing," *Wirel. Commun. Mob. Comput.*, vol. 2018, pp. 1–16, 2018, doi: 10.1155/2018/1934784.

[12] R. P. Noronha, "Diversity control in the hybridization GA-PSO with fuzzy adaptive inertial weight," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2021, pp. 1055–1062, doi: 10.1109/iciccs51141.2021.9432269.

[13] R. A. Rahman, S. Masrom, and N. Omar, "Tax avoidance detection based on machine learning of malaysian government-linked companies," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2S11, pp. 535–541, Nov. 2019, doi: 10.35940/ijrte.B1083.0982S1119.

[14] S. Masrom, R. A. Rahman, N. Baharun, and A. S. A. Rahman, "Automated machine learning with genetic programming on real dataset of tax avoidance classification problem," in *Proceedings of the 2020 9th International Conference on Educational and Information Technology*, Feb. 2020, pp. 139–143, doi: 10.1145/3383923.3383942.

[15] A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists*. O'Reilly Media, Inc., 2018.

[16] J. Lismont *et al.*, "Predicting tax avoidance by means of social network analytics," *Decis. Support Syst.*, vol. 108, pp. 13–24, Apr. 2018, doi: 10.1016/j.dss.2018.02.001.

[17] A. Hussain and Y. S. Muhammad, "Trade-off between exploration and exploitation with genetic algorithm using a novel selection operator," *Complex Intell. Syst.*, vol. 6, no. 1, pp. 1–14, Apr. 2019, doi: 10.1007/s40747-019-0102-7.

[18] J. Sato, T. Yamada, K. Ito, and T. Akashi, "Performance comparison of population-based meta-heuristic algorithms in affine template matching," *IEEJ Trans. Electr. Electron. Eng.*, vol. 16, no. 1, pp. 117–126, 2021, doi: https://doi.org/10.1002/tee.23274.

[19] Z. Beheshti and S. M. H. Shamsuddin, "A review of population-based meta-heuristic algorithms," *Int. J. Adv. Soft Comput. Appl.*, vol. 5, no. 1, 2013.

[20] S. S. Aote, M. M. Raghuwanshi, and L. G. Malik, "Improved particle swarm optimization based on natural flocking behavior," *Arab. J. Sci. Eng.*, vol. 41, no. 3, pp. 1067–1076, Mar. 2016, doi: 10.1007/s13369-015-1990-5.

[21] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimed. Tools Appl.*, vol. 80, no. 5, pp. 8091–8126, Feb. 2021, doi: 10.1007/s11042-020-10139-6.

[22] C. Blum and A. Roli, "Hybrid metaheuristics: an introduction," in *Hybrid Metaheuristics*, Springer Berlin Heidelberg, 2008, pp. 1–30.

[23] Z. Qin, F. Yu, Z. Shi, and Y. Wang, "Adaptive inertia weight particle swarm optimization," in *International conference on Artificial Intelligence and Soft Computing*, Springer Berlin Heidelberg, 2006, pp. 450–459.

[24] D. Chen and C. Zhao, "Particle swarm optimization with adaptive population size and its application," *Appl. Soft Comput.*, vol. 9, no. 1, pp. 39–48, Jan. 2009, doi: 10.1016/j.asoc.2008.03.001.

[25] A. Sarangi, S. Samal, and S. K. Sarangi, "Analysis of gaussian & cauchy mutations in modified particle swarm optimization algorithm," *019 5th Int. Conf. Adv. Comput. Commun. Syst.*, pp. 463–467, 2019.

BIOGRAPHIES OF AUTHORS

Associate Professor Ts. Dr Suraya Masrom is the head of the Machine Learning and Interactive Visualization (MaLIV) Research Group at Universiti Teknologi MARA (UiTM) Perak Branch. She received her Ph.D. in Information Technology and Quantitative Science from UiTM in 2015. She started her career in the information technology industry as an Associate Network Engineer at Ramgate Systems Sdn. Bhd (a subsidiary of DRB-HICOM) in June 1996 after receiving her bachelor's degree in computer science from Universiti Teknologi Malaysia (UTM) in Mac 1996. She started her career as a lecturer at UTM after receiving her master's degree in computer science from Universiti Putra Malaysia in 2001. She transferred to the Universiti Teknologi MARA (UiTM), Seri Iskandar, Perak, Malaysia, in 2004. She is an active researcher in the meta-heuristics search approach, machine learning, and educational technology. She can be contacted at email: suray078@uitm.edu.my.

Dr. Rahayu Abdul Rahman is an Associate Professor at the Faculty of Accountancy, UiTM. She received her PhD in Accounting from Massey University, Auckland, New Zealand in 2012. Her research interest surrounds areas, like financial reporting quality such as earnings management and accounting conservatism as well as financial leakages including financial reporting frauds and tax aggressiveness. She has published many research papers on machine learning and its application to corporate tax avoidance. She is currently one of the research members of the Machine Learning and Interactive Visualization Research Group at the UiTM Perak Branch. She can be contacted at email: rahay916@uitm.edu.my.

Dr. Masurah Mohamad is currently a Senior Lecturer at Universiti Teknologi MARA Perak Branch, Tapah Campus (UiTM Perak) which is an educational institute established by the Ministry of Higher Education Malaysia for 15 years in the Department of Computer Science. She received her Ph.D. in Computer Science from UTM in 2021. Before joining Universiti Teknologi MARA, she served with Management and Science University (MSU) for 1 year. She has received several research grants such as the Fundamental Research Grant (FRGS) funded by the Ministry of Higher Education in 2012 and 2021, and the Lestari Research Grant funded by Universiti Teknologi MARA (UiTM) in 2019 and other internal and external research grants (2009-2020). Currently, she is serving on the Editorial Boards of Mathematical Sciences and Informatics Journal (MIJ) under UiTM Press Publications, Malaysia. She is a Publication chair for the 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS2019) and the 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS2021). She also has contributed to several conferences organized by UiTM Perak Branch as secretariat committee (2010, 2013, and 2015) and reviewed several articles for several journals and conferences. Her research interests include data sciences and analytics, data mining and information retrievals, Artificial Intelligence and machine learning, recommender systems, soft computing, and data visualization. She can be contacted at email: masur480@uitm.edu.my.

Ts. Abdullah Sani Abd Rahman obtained his first degree in Informatique majoring in Industrial Systems from the University of La Rochelle, France in 1995. He received a master's degree from Universiti Putra Malaysia in Computer Science, with a specialization in Distributed Computing. Currently, he is a lecturer at the Universiti Teknologi PETRONAS, Malaysia and a member of the Institute of Autonomous System at the same university. His research interests are cybersecurity, data analytics and machine learning. He is also a registered Professional Technologist. He can be contacted at email: sani.arahman@utp.edu.my.

Dr. Norhayati Baharun is an Associate Professor of Statistics, Universiti Teknologi MARA Perak Branch, Tapah Campus. She received her PhD in Statistics Education from the University of Wollongong Australia in 2012. Her career started as an academic from January 2000 to date at the Universiti Teknologi MARA that specialized in statistics. Other academic qualifications include both Master Degree and Bachelor Degree in Statistics from Universiti Sains Malaysia and Diploma in Statistics from Institute Teknologi MARA. Among her recent academic achievements include twelve on-going and completed research grants (local and international), four completed supervision of postgraduate studies, fifteen indexed journal publications, two academic and policy books, twenty-six refereed conference proceedings and book chapter publications, a recipient of 2013 UiTM Academic Award on Teaching, and fourteen innovation projects with two registered Intellectual Property Rights by RIBU, UiTM. She is also a certified Professional Technologist (Ts.) (Information & Computing Technology) of the Malaysia Board of Technologist (MBOT), a Fellow Member of the Royal Statistical Society (RSS), London, United Kingdom, a Professional Member of the Association for Computing Machinery (ACM), New York, USA, and a Certified Neuro Linguistic Program (NLP) Coach of the Malaysia Neuro Linguistic Program Academy. Her research interests continue with current postgraduate students under her supervision in the area of decision science now expanding to a machine learning application. She can be contacted at email: norha603@uitm.edu.my.