Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 10, No. 4, Desember 2020

ISSN 0216 – 0544
e-ISSN 2301– 6914

APPLICATION OF HYBRID GA-PSO TO IMPROVE THE PERFORMANCE OF DECISION TREE C5.0

[a]Achmad Zain Nur, [b]Hadi Suyono, [c]Muhammad Aswin

[a,b,c] Departement of Electrical Engineering Brawijaya University, Malang, Indonesia
E-mail: [a] achmadnur.zn@gmail.com, [b] hadis@ub.ac.id, [c] maswin@ub.ac.id

Abstract

Data mining is a data extraction process with large dimensions and information with the aim of obtaining information as knowledge to make decisions. Problems in the data mining process often occur in high-dimensional data processing. The solution to handling problems in high-dimensional data is to apply the hybrid genetic algorithm and particle swarm optimization (HGAPSO) method to improve the performance of the C5.0 decision tree classification model to make decisions quickly, precisely and accurately on classification data. In this study, there were 3 datasets sourced from the University of California, Irvine (UCI) machine learning repositories, namely lymphography, vehicle, and wine. The HGAPSO algorithm combined with the C5.0 decision tree testing method has the optimal accuracy for processing high-dimensional data. The lymphography and vehicle data obtained an accuracy of 83.78% and 71.54%. The wine dataset has an accuracy of 0.56% lower than the conventional method because the data dimensions are smaller than the lymphography and vehicle dataset.

Key words: Data Mining, Decision Tree, Hybrid GA – PSO, High Dimensional Data.

INTRODUCTION

Data mining is an extraction process of data with large dimension to obtain information as knowledge to make decisions. Machine learning is part of data mining to help find data patterns automatically. Method that requently used in data mining is decision tree C5.0 [1].

In the mining process, the data used sometimes has problems that can interfere with the results of the mining process. Among them are missing values, redundant data, outliers, or data formats that are incompatible with the system. This problem often occurs in high dimensional data or data that has a large dimensional size. Some methods that are often used to solve problems in high dimensional data are the genetic algorithm (GA) and the particle swarm optimization (PSO) [2].

Various methods have been developed in several studies to solve problems in high dimensional data. First research is about Feature Selection for Varying Coefficient Models in Ultrahigh-Dimensional Covariates. The performance result in this study was 95% [3]. Another research is Model-Free Feature Screening in Ultrahigh Dimensional Discriminant Analysis. The performance result of this method is 94.30% [4]. Other research uses hybrid genetic algorithm and particle swarm optimization methods to solve bi-level linear programming problems. In this study, it is able to minimize the error rate to 0.0094 and show that the optimization results using only GA or PSO are no better than the hybrid method [5]. Research that has been carried out by conventional optimization techniques is considered not optimal in dealing with feature selection problems on high dimensional data, therefore it is necessary to optimize feature selection in the decision tree C5.0 method by maximizing the application of the model in preprocessing [6].

This research is a solution for dealing with problems in high dimensional data by using the hybrid genetic algorithm and particle swarm optimization methods to improve the performance of the decision tree classification model C5.0 to make decisions quickly, precisely and accurately on classification data.

MATERIAL AND METHODS

Data Mining

Data mining is also called knowledge discovery in database (KDD). Data mining has three main points which are shown in figure 1.

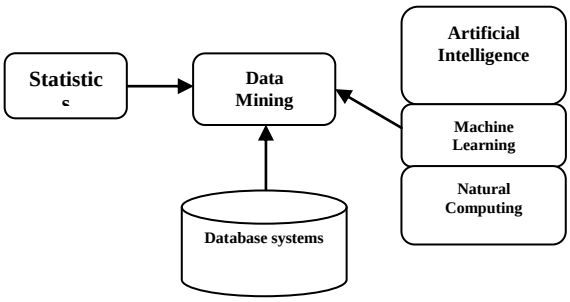

Fig 1. Data mining root

Based on Figure 1, statistics are the most important points in data mining. Statistics are used to identify systematic relationships between different variables, when there is not enough information on those variables. Artificial Intelligence (AI) contributes to data processing techniques, based on human reasoning models for data mining development [2].

Closely related to AI, Machine Learning (ML) is very important in data mining development. ML uses techniques that allow computers to learn by 'training'. In this context also consider Natural Computing (NC) as a solid additional root for data mining. Databases Systems (DBS) provide information which is then processed using data processing methods [2].

High Dimensional Data

High dimensional data can help machine learning models to learn more rules and better generalize new data [4]. However, adding low-quality data and reckless input features may create too much noise and can slow down the training algorithm. A number of techniques for data dimension reduction are available to estimate how informative each column is and, if necessary, to filter it from the dataset [6]. Here are some data-dimensionality reduction techniques [7].

1. Ratio of missing value (calculating the ratio of empty values in data)
2. Low variance in the column value (trim low variance in the column)

3. High correlation between two columns (making iterations to calculate the correlation between 2 columns)
4. Principal component analysis (analyzing the main components of the data)
5. Candidates and split columns in a random forest (separating columns and applying a random forest model to the data)
6. Backward feature elimination
7. Forward feature construction (construct data attributes forward)

Selection of Genetic Algorithm Attributes

Genetic algorithm (GA) is an optimization and search technique based on the principles of genetics. Genetic algorithms mainly consist of three operators: selection, crossover, and mutation [8]. The flowchart of the GA attribute selection method is shown in Figure 2.

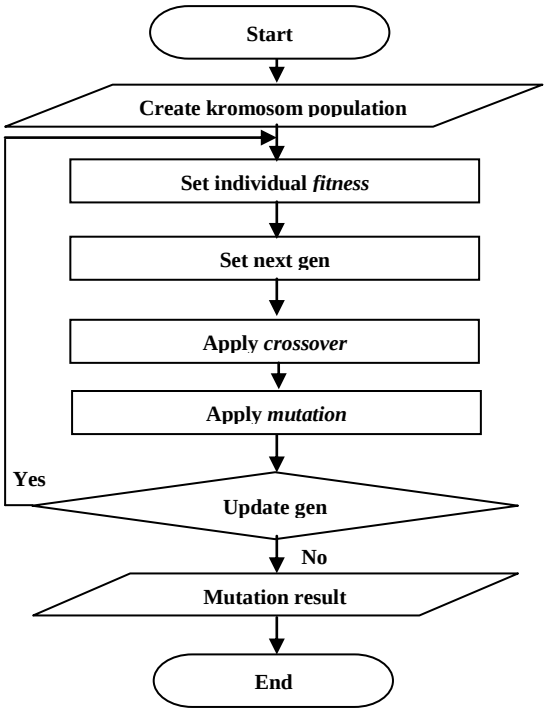

Fig 2. Genetic algorithm flowchart
(Source: H. Hachimi, R. Ellaia, A. Elhami 2012)

In Figure 2, the fitness function in GA is a simple function, assigning ratings to individual attributes at the bottom of the correlation coefficient. It can be said that the lower the correlation, the higher the fitness value of the attribute. The relative fitness function is as follows [9]:

$$P[i] = \frac{f[i]}{\sum f} \quad (1)$$

f = fitness value
∑f = the total fitness value of all chromosomes
i = the iteration of chromosome
P = relative probability

Then the results of calculation 1 will be used as a value material to find the cumulative fitness. The calculation of the cumulative fitness is carried out as much as the number of the population owned, so that there will be groups that have a certain value distance in each group. Cumulative fitness is calculated as follows [9]:

$$C[i] = C[i-1] + P[i] \quad (2)$$

C = cumulative fitness
i = the iteration of chromosome
P = relative probability

Then form a random number between 0 and 1 and check the position of the random number on the given relative probability.

Particle Swarm Optimization Attribute Selection

The PSO method is made on the basis of the movement activities and behavior of a group of fish and flocks of birds in behavior such as looking for prey, which was first proposed by James Kennedy and Russell C. Eberhart in 1995. PSO consists of a group of particles looking for the best position, which is the best position for optimization problem in feature space [11]. The schematic of the particle swarm optimization attribute selection method is shown in Figure 3.

Based on Figure 3, the initialization of the PSO algorithm begins by assigning a random initial position of the particle (solution) and then searching for the optimal value by updating its position. As explained above, each iteration of each particle updates its position according to the two best values, namely the best solution that has been obtained by each particle (pbest) and the best solution in the population (gbest) [10].

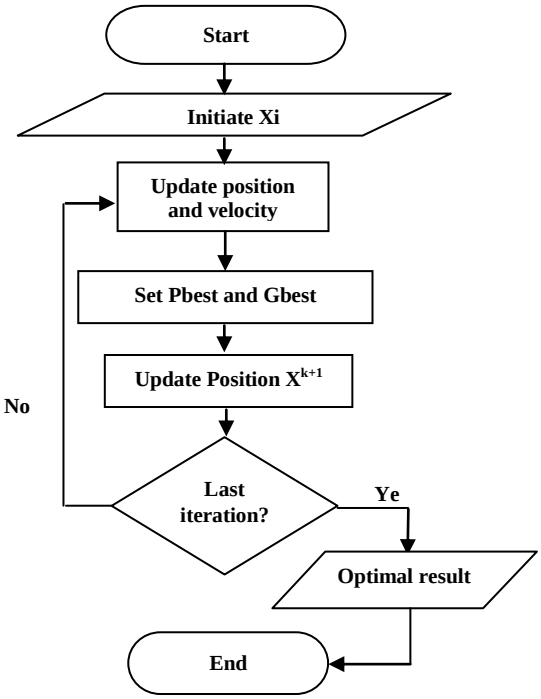

Fig 3. Particle Swarm Optimization Flowchart (Source: H. Hachimi, R. Ellaia, A. Elhami 2012)

The PSO method can be applied using equations (3) and (4) shown below to update the velocity and position of each particle.

$V_{id}^{k+1} = W.\ V_{id}^{k} + c1.\ rand1.\ (P_{id} - X_{id}) + c2.\ rand2.\ (G_{id} - X_{id})$ (3)

The equation for determining the attribute position of i on d dimension as follows.

$$X_{id}^{k+1} = X_{id}^{k} + V_{id}^{k+1} \quad (4)$$

Vid = the iteration individual velocity component in d dimension

Xid = individual position i in d dimension

ω = parameter of inertia weight

$c_1 c_2$ = acceleration constant (learning rate), the value is between 0 to 1

$rand_{1,2}$ = random parameter between 0 to 1

P_{id} = Pbest (local best) individual i on d dimensions

G_{id} = Gbest (global best) in d dimensions

Vid is the velocity of particle in the iteration k, and Xid is the solution (position) of the particle in k iteration. c1, c2 are positive constants, and rand1, rand2 are two random variables using uniform distribution between 0 to 1. W is the inertion weight which shows the effect of changing the velocity from the old vector to the new vector [11].

HGAPSO Attribute Selection

The method of hybrid genetic algorithm and Particle Swarm Optimization is a method that is carried out in two phases to produce a new population. The hybrid model is performed by selecting N randomly generated individuals. The new individual can be considered a chromosome in the case of GA or called a particle in the case of PSO. N individuals are sorted by fitness, and the best N individuals are entered into the GA model to make N individuals new by crossover [12].

The crossover operator in GA is applied using the concept of a linear combination of two vectors, which shows two individuals in the GA algorithm with 100% crossover probability. The random mutations generated by the GA algorithm will be replaced by the PSO method [6]. The procedure for adjusting the N particles in the PSO method involves selecting particles globally, selecting particles from the best population, and then updating the velocity values. The best global particle population is determined according to the fitness value that has been sorted [13].

The following is the pseudo code notation of the hybrid genetic algorithm and particle swarm optimization [12].

1. GA Method
Generate a population of size 4N for an N-dimensional problem
Repeat
For i=1 to N do
Evaluate the fitness of each of the N individuals
Rank them on the basis of the fitness values
Selection (apply to the top 2N individuals and create another 2N individuals)
If F(x1) > Fbest so
Fbest = F(x1)
End for
100% crossover
For the N best individuals, apply two-parent crossover to update
The N best individuals
End for
For
GA-Mutation with 20% mutation probability in 2N best chromosomes according to the equation below:

$$Xi = xi + rand \times N(0,1) \quad (5)$$

End for

2. PSO Method
Apply PSO operators (velocity and position updates)
For the updating the N individuals with worst fitness
Update the particles velocity and position

$V_{id}^{new} = W.\ V_{id}^{old} + c1.\ rand1.\ (P_{id} - X_{id}) + c2.\ rand2.\ (G_{id} - X_{id})$ (6)

$X_{id}^{new} = X_{id}^{old} + V_{id}^{new}$ (7)

With c1 = c2 and w = [0.5 + rand/2.0]

Until the termination criterion is reached

Equation 6 illustrates that the new velocity of each particle is updated with the previous velocity (V_{id}), the best location in the particle

population (P_{id}) and the best global location (P_{gd}). The velocity particles in each dimension are sandwiched using V_{max} which is arranged into certain blocks of the search space for each dimension i. Equation 7 shows that each particle (X_{id}) is updated during the search for the best solution [12].

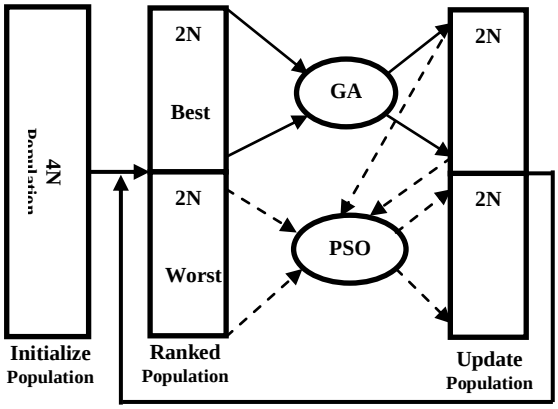

Fig 4.Scheme of Hybrid Genetic Algorithm Particle Swarm Optimization Method
(Source: Yi-Tung Kao, Erwie Zahara 2008)

Figure 4 shows a schematic HGAPSO method by setting the initial population as 4N. The dotted line shows the GA process, the dotted line shows the PSO process.

Algorithm C5.0

Algoritm a C5 .0 is a refinement of the previous algorithm, namely C4.5 and ID3. Compared with C4.5, C5.0 algorithms faster and more effective to generate decision-making tree. Information gain is a separation criterion that uses entropy measurements. To get information gain from an attribute, it takes the entropy of the whole class or Entropy (S) Entropy (S) is the estimated number of bits needed to be able to extract a class from a number of random data in the sample space. Mathematically, entropy is formulated as follows [14].

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i)) \qquad (8)$$

After getting the entropy value, then look for the information gain value. Information gain is used to measure the effectiveness of attribute characteristics in classify classes. Equation 2.6 is used to calculate the information gain as follows [15] :

$$IG(S, A_i) = H(S) - \sum_{\alpha \in A_i}^{n} \frac{|S_\alpha|}{|S|} H(S_\alpha) \qquad (9)$$

H = Entropy
S = Case set
A = Attribute
n = Number of partitions attribute A
$|S_\alpha|$ = number of cases on partition i
| S | = Number of cases in S.
p_i = Proportion of S_i to S

Gain (A) is the expected reduction in entropy caused by knowledge of the value of attribute A. The algorithm calculates the information gain for each attribute. The attribute with the greatest gain value is chosen as the attribute test (node root). A node is created and labeled with attributes, branches are created for each attribute value [14] .

Research Methods

This test uses Microsoft Excel 2013 to process the dataset and Rapid Miner 5.3.0 software to design and analyze the results of the calculation method. The research was conducted on the data collection phase, the initial data processing, application HGAPSO method, the application of the classification model, testing, and validation.

The data collection stage is the selection of data sets from the UCI Machine Learning Repository. Then the initial data processing is carried out on the selected data set. Initial processing includes data cleaning processes such as replace missing values, removing duplication values, noise and outliers. After the data were normalized, the HGAPSO method was applied.

The next step is to apply the decision tree C5.0 classification model and test it. The classification results are then evaluated and validate the results of the research performance. The data used in this study came from open source data, namely the UCI Machine Learning Repository. The research data used is shown in table 1.

Table 1. Research Dataset

No.	Dataset	Data Record	Atrributes	Number of Class
1	Lymphography	148	18	4
2	Vehicle	946	18	4
3	Wine	178	13	3

(Source: UCI Machine Learning Repository)

In Table 1, the dataset lymphography is data that utilizing x-ray technology to view the lymphatic circulation and gland lymph in the diagnosis of disease. Dataset vehicle is data about what types of vehicles based angel view different images. Dataset wine is the analysis of wines which were planted in the area the same in the field of chemical research . The dataset is selected based on different attribute dimensions to test the effectiveness of the HGAPSO algorithm.

Initial processing of the dataset is carried out to obtain quality data, some of the techniques used are Data Selection, Min-Max Normalization, and Attribute Selection .

The classification process is shown in Figure 5.

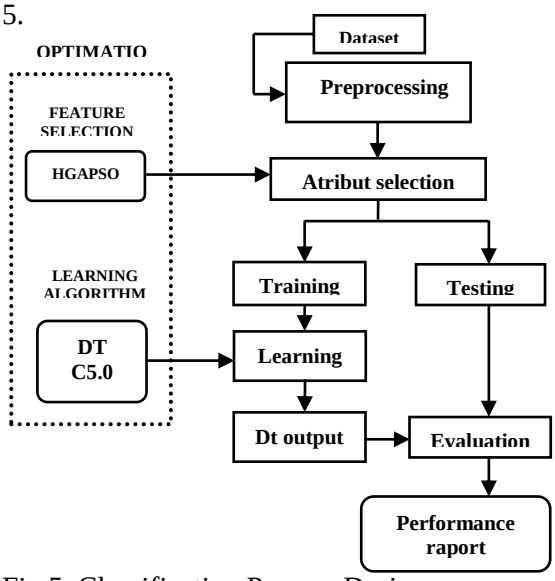

Fig 5. Classification Process Design
(Source: Research Result)

HGAPSO Concept

The application of the hybrid method starts with the GA method then continues with the PSO method. In the hybrid method, the population generation process is limited to 4N chromosomes to produce the best 2N individuals and produce 2N new individuals. The application of PSO is carried out to process the worst 2N individuals in the population generation. The initial steps in the GA method are population initialization, evaluation function, selection process, crossover, and mutation. The following HGAPSO algorithm steps:

1. Initialization: Determine the initial population generation of 4N
2. Evaluation and ranking: Evaluate the fitness value of each individual 4N and sort by the best value using the equation:

$$D = \sum_{i=0}^{n} w_i \times s_i) \quad (10)$$

n = number of chromosomes
Si = the binary value of the iteration gene
W = weight in the iteration gene
I = iteration of chromosome index

3. GA method : Apply the GA operator (crossover and mutation) to the best 2N and create 2N new individuals.
   a. (Selection): from each population, select the best 2N individuals based on the fitness value
   b. (100% Crossover): Apply two-parent crossover to update the best individual 2N
   c. (20% Mutation): Applying a mutation with a 20% probability of mutation to the updated 2N chromosomes
4. PSO method: Apply operator PSO (velocity and position updates) to 2N new individuals who are passed on the GA stage. Update particle velocities and positions in the updated concept of cost-sensitive decision C5.0

Testing Method

The multiclass confusion matrix is a method used to measure the performance of a classification method [19] . In this study, the parameters used to measure the classification performance are accuracy, precision, recall, and F-measure. D ata are classified into several class as shown in Table 2 below:

Table 2. Confusion Matrix

		Predicted Class		
		1	**2**	**3**
Actual Class	**1**	$N_{(1,1)}$	$N_{(1,2)}$	$N_{(1,3)}$
	2	$N_{(2,1)}$	$N_{(2,2)}$	$N_{(2,3)}$
	3	$N_{(3,1)}$	$N_{(3,2)}$	$N_{(3,3)}$

(Source: Akbar, Yudhistira and Cholissodin, 2014)

Based on table 2, a confusion matrix table can be used to determine performance parameters using the following conditions:

Accuracy = (TP+TN) /Ntotal
Precision = TN / (TN + FP)
Recall = TP / (TP + FN)
F-Measure = 2 x ((Recall x Precision) / ((Recall + Precision))

1. True Postive (TP) shows that the class generated by the classification prediction is positive and the actual class is positive.
2. True Negative (TN) indicates that the class resulting from the classification

prediction is negative and the actual class is negative.

3. False Positive (FP) indicates that the class generated from the classification prediction is negative and the actual class is positive.
4. False Negative (FN) indicates that the resulting class from the prediction of the classification is positive and the actual class is negative.

RESULT AND DISCUSSION

Data processing begins by setting an initial population of 4N attributes . The best 2N attribute is processed using the GA algorithm to produce a new 2N population mutation. The worst 2N atibut is processed using the PSO algorithm to produce optimal particles . The new 2N fertilization and optimal particles are then reprocessed by determining the next 4N population generation , to produce a new optimal population. Then the new population was tested using the DT C5.0 method. The accuracy performance result is shown in table 3.

Table 3. Parameter Value of Acuracy C5.0

Lymphography	**76,35 %**	**83,11 %**	**83,11 %**	**83,78 %**
Vehicle	**67,82 %**	**71,28 %**	**68,88 %**	**71,54 %**
Wine	**91,57 %**	**97,19 %**	**96,63 %**	**96,63 %**

(Source: Calculation Results)

Based on Table 3, classification results in green indicate the best percentage, red indicates the worst and yellow indicates intermediate. The test results obtained by the accuracy value used to measure the performance of the method in classifying the dataset . The performance of the C5.0 algosrithm is small compared to the optimized performance of C5.0. Algorithm C5.0 which has been optimized by the use of PSO on the dataset lymphography, vehicle, wine can be increased respectively 83,11%, 71.28%, 97.19%. While the results of the C5.0 algorithm using GA on the sympographic dataset , vehicle, wine were able to increase the accuracy value of 83.11%, 68.88%, 96.63%, respectively.

The test results using the GA PSO hybrid method on the C5.0 decision tree using the hybrid method on the sympography and vehicle dataset can increase the accuracy higher than the conventional method, namely at 83.78% and 71.54%. While in the dataset wine have accuracy values were slightly lower than optimization PSO, namely 96 , 63 %. The results of the DT C5.0 + HGAPSO test are shown in table 4.

Table 4. Test Results DT C5.0 + HGAPSO

Dataset	Accuracy	Recall	Precision	F-measure
Lymphography	83,78%	83,78%	94,59%	88,86%
Vehicle	71,54%	71,54%	90,51%	79,92%
Wine	96,63%	96,63%	98,31%	97,46%

(Source: Calculation Results)

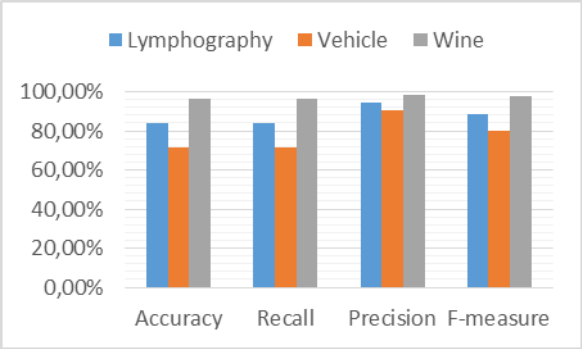

Fig 6. DT C5.0 + HGAPSO Result

(Source: Calculation Results)

Dataset	C5.0	C5.0 + PSO	C5.0 + GA	C5.0 + HGAPSO

Based on Table 4 and Figure 6 , the results of testing the lymphography, vehicle, and wine dataset obtained optimal performance results from 4 test parameters, namely accuracy, recall, precision, and F-measure..

CONCLUSION

Based on the test results using the HGAPSO method on the DT C5.0 using the hybrid method on the lympography dataset and the vehicle, it can increase the accuracy higher than the conventional method at 83 , 78 % and 71.54%. Meanwhile, the wine dataset has a slightly decreased accuracy value compared to the PSO optimization of 96.63%. This is because the dimensions of the wine dataset are smaller than the lymphography and vehicle dataset. The GA-PSO hybrid method can be quite effective in improving classification performance on high-dimensional data.

REFERENCES

[1] Laksmi B. N., Indumathi T. S., Nandini R., "A study on C.5 Decision Tree Classification Algorithm for Risk Predictions during Pregnancy." *ICETEST,* vol 24, 2015, pp. 1542-1549.

[2] Larose D, T, "Discovering knowledge in data: an introduction to data mining." *Jhon Wiley & Sons Inc*., 2005.

[3] J. Liu, R. Li, and R. Wu, "Feature Selection for Varying Coefficient Models With Ultrahigh-Dimensional Covariates." *Journal of the American Statistical Association,* vol. 109, no. 505, 2014, pp. 266-274.

[4] Cui, R. Li, and W. Zhong, "Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis." *Journal of the American Statistical Association,* 2014.

[5] R.J. Kuo, Y.S. Han, "A hybrid of genetic algorithm and particle swarm optimization for solving bi-level linear programming problem – A case study on supply chain model." *Applied Mathematical Modelling,* vol. 35, 2011, pp. 3905-3917.

[6] S. Malavika, K. Selvam, "Reduction of Dimensionality for High Dimensional Data using Correlation Measures." *Global Journal of Pure and Applied Mathematics (GJPAM),* vol. 11, no. 1, 2015, pp. 107-111.

[7] Yin L, Y. Ge, K. Xiao, X. Wang, and X. Quan. "Feature selection for high-dimensional imbalanced data." *Neurocomputing,* vol. 105, 2013: pp. 3–11.

[8] R. Tiwari, M. Pratap Singh, "Correlation-based Attribute Selection using Genetic Algorithm." *International Journal of Computer Applications,* vol. 4, no. 8, 2010, pp. 28–34.

[9] I., Wang, M. Q. Li. "The application of data mining technology based on genetic algorithm." *Journal of Nanchang University,* vol. 1, no. A27, 2007, pp. 81-84.

[10] Wei S, Y.K. Ching, C.S. Chieh and L.Z. Jung. "Particle Swarm Optimization for Parameter Determination and Feature Selection of Support Vector Machines." *ScienceDirect: Expert System With Aplications,* 2008: pp.1817- 1824.

[11] Lin S. W and Ying K. C." Particle swarm optimization for parameter determination and feature selection of support vector machines." *Expert Systems with Applications*. 2008: pp. 1817–1824.

[12] Hachimi, R. Ellaia, and A. Elhami, "A New Hybrid Genetic Algorithm and Particle Swarm Optimization." *Key Engineering Materials,* vol. 498, 2012, pp. 115-125.

[13] Ahmed F. Ali, Mohamed A. Tawhid, "A Hybrid Particle Swarm Optimization And Genetic Algorithm With Population Partitioning For Large Scale Optimization Problems." *Ain Shams Engineering Journal,* 2016.

[14] Polat K. and S. Gunes. "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-classclassification problems." *Expert Systems with Applications*, vol.36, 2009. pp: 1587–1592.

[15] Ahmed F. Ali, Mohamed A. Tawhid, "A Hybrid Particle Swarm Optimization And Genetic Algorithm With Population Partitioning For Large Scale Optimization Problems." *Ain Shams Engineering Journal,* 2016.

[16] J. Arunadevi, M. Josephine Ninthya, "Comparison of Feature Selection Strategies for Classification using Rapid Miner." *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 4, Issue 7, 2016, pp 13556–13563.

[17] Wang A, R. Li and Z. Chen. "Partition cost-sensitive CART based on customer value for Telecom customer churn prediction." *Proceedings of the 36th Chinese Control Conference.*, 2017.

[18] John W, "Data Mining, Modelling and Management". *International Journal*, 2019, ISSN online 1759-1171, ISSN print 1759-1163