

PAPER • OPEN ACCESS

Feature Selection using Simulated Annealing with Optimal Neighborhood Approach

To cite this article: A Syaiful *et al* 2021 *J. Phys.: Conf. Ser.* **1752** 012030

View the [article online](#) for updates and enhancements.



240th ECS Meeting

Digital Meeting, Oct 10-14, 2021

**Register early and save
up to 20% on registration costs**

Early registration deadline Sep 13

REGISTER NOW



Feature Selection using Simulated Annealing with Optimal Neighborhood Approach

A Syaiful^{1*}, B Sartono¹, F M Afendi¹, R Anisa¹, and A Salim²

¹ Department of Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia

² Department of Mathematics and Statistic, La Trobe University, Melbourne 3086, Australia

*Email: bagusco@apps.ipb.ac.id

Abstract. The one of the metaheuristic approaches that can be used was simulated annealing (SA) algorithm which inspired by annealing metallurgical process. This algorithm shows advantages in finding global optimum of given function which will be used in feature selection. In this study, we will try to combine the neighborhood size and limited approach by using data simulation comparing between two function which is Akaike Index Criterion (AIC) function and Bayesian Index Criterion (BIC) function. The result of this experiment shows that the selected variables using optimal neighborhood size and limit the selected variable provide the result of goodness model around 98% of accuracy and specificity and 94% of sensitivity compared with simulated annealing algorithms without any modification using both AIC function and BIC function, and in the simulation also shows that BIC function give better result than AIC function.

Keywords: Simulated Annealing, Neighborhood Size, AIC, BIC

1. Introduction

In recent developments of science and technology have led to the development of vast amounts of data. A lot of this data can be useful for researcher or data scientist to get important information using the right method. At some point, it would encounter on regression problems where the number of features or covariates p exceeds the sample size n . This type of data can also be called as high dimensional regression where we have big size of data. In high dimensional regression case, we can generally find several cases like one-level factors, two-level factors, or even three-level factors. This case can be used depend on how we define our case of regression.

In high dimensional regression, of course we are not going to use all variable because not all those variables would have significant impact to our data and give some information. So instead of using the all variables, we need to select only certain variables which have impact or significant to the data. We could do feature selection because we need to priority which variables would be necessary to used. Selecting variable also can mean to finding the simplicity of modelling when we using only certain variables. Beside both reason, we also need to make our modelling become more effective and efficient by using not all data into our model.

There are several techniques that we can use to select significant variables on high dimensional regressions. Several methods are included forward selection, backward selection, stepwise selection, best subset, ridge and lasso, and so on which general method to find variable in the best way. There also method that we can use to select variable are optimization or metaheuristic method which more recent introduce into feature selection. Genetic algorithm, simulated annealing, ant colony



optimization, particle swarm optimization, and so on are such optimization or metaheuristic method that have been developed to select variable.

A metaheuristic is a high-level problem-independent algorithmic framework that provides a set of guidelines or strategies to develop heuristic optimization algorithms. The term is also used to refer to a problem-specific implementation of a heuristic optimization algorithm according to the guidelines expressed in such a framework [1]. This method is able to produce a good solution with huge potential to capture what researchers and data scientist need, but even so it does not guarantee that the resulting solution is the best solution but surely effective to help them.

One of metaheuristic method that have been developed and can be used in feature selection and also would be used in this paper was simulated annealing. Simulated annealing itself is one of the optimization methods which consist of combination of probabilistic theory and statistical mechanics inspired by the annealing metallurgical process. This algorithm shows advantages in finding global optimum of given function which will be used in feature selection. Akaike Index Criterion (AIC) and Bayesian Index Criterion (BIC) are function that we will use to optimize to finding significant variable that can be more effective. Besides using both functions, increasing neighborhood size also can be effective to increasing more faster way to select variables. By using this, hopefully we can get better results in feature selection.

The remainder of this the paper is organized as follows. In methods session, we discuss about proposed method and simulation experiment that we will using in this paper included Akaike Index Criterion (AIC) function and Bayesian Index Criterion (BIC) function which would be our cost function that we trying to optimize, and to evaluate goodness of model in this selection by using measures like accuracy, sensitivity, and specificity. This goodness of model would show us how good our selection. We also show the details of how the methods are implemented and the performance in result and discussion session. Last but not least chapter, we conclude our discussion in conclusion session.

2. Method

2.1. Regression model

The high dimensional regression in this paper would be focus in one-level factor which the format more likely simple linear regression. The simple linear regression method models which the relationship between two variables with general shapes [2]:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \varepsilon \quad (1)$$

or

$$y = x\beta + \varepsilon \quad (2)$$

where y is the response variable, $\beta_0 \dots \beta_n$ is the regression coefficient and $x_1 \dots x_n$ is the explanatory variable, and ε is an error in this model.

2.2. Proposed Method

Simulated annealing (SA) is a metaheuristic optimization method whose main principle is inspired by the metallurgical field which consists of a combination of probabilistic theory and statistical mechanics by adopting the principle of annealing. In metallurgy, the annealing method is the process of hardening the material which is done by heating the metal to its melting point, maintaining it at a certain temperature before then cooling it down. Initialized with high temperature parameters, it does a global random search from one neighbour to another neighbour. In the second stage, the temperature decreases progressively and the search becomes local.

Simulated annealing algorithm is generally used if the search space is discrete. There are four components needed to perform this simulated annealing algorithm [3]. The four components are:

- a) Configuration: a model that represents all possible solutions that can be taken and will then be used to find the optimal solution. This represent the possible problem solution over which we will search for good answer.
- b) Move Set: a set of allowable moves that permit us to reach all feasible configurations and one that is easy to compute. These moves are the computations we must perform to move from configuration to configuration as annealing process.

- c) Objective Function (OF): to measure how good any given placement configuration is.
- d) Cooling Schedule: to anneal the problem from a random solution to a good, frozen, placement. Specifically, we need a starting hot temperature (or heuristic for determining a starting temperature for current problem) and rules to determine when the current temperature should be lowered, and when annealing should be terminated.

The feature selection algorithm based on simulated annealing proceeds as follows [4]. Firstly, an initial solution is selected randomly and it is assumed to be the optimal solution. Subsequently, the cost of the initial solution is computed using the cost function described previously. While temperature T does not satisfy the termination condition, a neighboring solution of the current optimal solution is selected and its cost is calculated. If the cost of the newly selected neighboring solution is less than or equal to that of the current optimal solution, the current optimal solution is replaced with a newly selected neighbor solution. If the cost of the neighboring solution is greater than the current optimal solution, a random value q is chosen in the range of $(0, 1)$. In this case, the replacement of the optimal solution is permitted only if a random value q is less than $e^{-\frac{\text{Cost}(Vn) - \text{Cost}(Vb)}{T}}$. After temperature T is reduced following (6), these processes are continued until T satisfies the termination condition.

A pseudo-code for the feature selection algorithm based on simulated annealing is suggested as follows. The cost of this simulated annealing algorithms would be using function of Akaike Index Criterion (AIC) and Bayesian Index Criterion (BIC). The value of the Akaike Index Criterion (AIC) fitness criterion is an estimate of the relative quality of the statistical model for a given set of data. AIC function values are calculated using the formula [5]:

Table 1. Basic simulated annealing algorithm to feature selection

Algorithm: Feature selection algorithm based on simulated annealing
Input: Training data set
Output: Combination of features: vb
1. vb $\leftarrow$ Null; // final solution
2. T $\leftarrow$ 100000;
3. r $\leftarrow$ 0.9;
4. Generate an initial solution, vi;
5. vb $\leftarrow$ vi;
6. Calculate the cost of initial solution, Cost(vb);
7. while (T > 0.001) do
8. begin:
9. Randomly select a neighbor solution, vn, of vb which have one bit different from vb;
10. if (Cost(vb) = Cost(vn)):
11. vb $\leftarrow$ vn;
13. else:
14. Generate a random number q uniformly in the range (0, 1);
15. if ($q < e^{-\frac{\text{Cost}(Vn) - \text{Cost}(Vb)}{T}}$)
16. vb $\leftarrow$ vn;
17. T $\leftarrow$ r $\times$ T
18. end // for while loop

$$\text{AIC} = -2.\ln(L) + 2k \quad (3)$$

where $\ln(L)$ is the maximum Likelihood of a model with k parameters. The smaller the AIC function value, the better the data set obtained.

While value of the Bayesian Index Criterion (BIC) fitness criterion is a criterion for model selection among a finite set of models. BIC function values are calculated using the formula [2]:

$$\text{BIC} = -2.\ln(L) + \ln(n).k \quad (4)$$

where $\ln(L)$ is the maximum Likelihood of a model with k parameters with n of number observation. The smaller the BIC function value, the better the data set obtained

The neighbor in our simulated annealing was not totally random, but more likely only have one different variables from original neighbor. So suppose we have original neighbor $X = (X_1, \dots, X_n)$ where X_i would have either 0 or 1, and we like to find the neighbor from this. Let said we would have $X^k = (X_1, \dots, X_n)$ where randomly select one variable from X , let say X_k and changing this variable from 0 to 1 or from 1 to 0. The size of this neighbor can also be increasing to get effective result. So, suppose that each iteration we only select one neighbor and compare the result with current or original neighbor. Now, we can increase this neighbor into K neighbor and find best result from this neighbor and compare the result with current or original neighbor.

We also like to evaluate goodness of model by using measures like accuracy, sensitivity, and specificity where the formula would be [6]:

	+R	-R		+R	-R	
+P	True Positive	False Positive		A	B	A+B
Figure 1. Confusion matrix						

$$\text{Accuracy} = \frac{A+D}{N} \quad (5)$$

$$\text{Sensitivity} = \frac{A}{A+C} \quad (6)$$

$$\text{Specificity} = \frac{D}{B+D} \quad (7)$$

Accuracy refers to the closeness of a measured value to a standard or known value, Sensitivity refers to closeness of true positive to all known positive value, and specificity refers to closeness of true negative to all known negative value.

2.3. Simulation Experiments

The data used in this study are simulation data with the following conditions:

- Generate 100 random variables with data of 100 observations that are spread uniformly between 1 to 100 with the first 10 variables as explanatory variables with the same regression coefficient values and the response variables are calculated from the 10 explanatory variables by adding them up plus the error values that spread normally with a mean value of 0 and range 0.1
- Select the simulated annealing process variable with various temperature measurements to see whether the temperature change affects the selection of the variable
- Do the same process with point b for one temperature and increase the size of the neighbors in the selection of the variable to see whether changes in the size of the neighbors also affects
- Do the same process with point c and add the selected variable selection p-value to see if the use of p-value affects the selection of variables
- Do the same process with point c and limit the selected variable to a certain maximum value to see whether the restrictions of the selected variable affect the selection of the variable.
- Repeat each process (point b to e) to 100 times to see the consistency of the selected variable
- Do the process f by using the cost value of AIC and BIC functions
- Calculate the measurement of accuracy, sensitivity and specificity to measure the level of goodness of the model

3. Results and Discussion

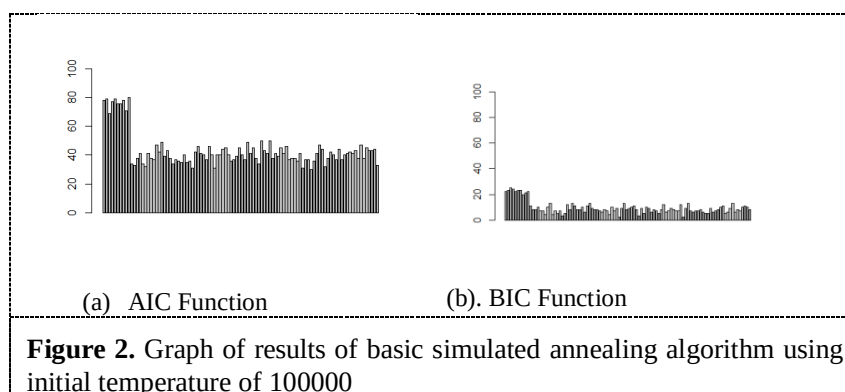
3.1. Basic of simulated annealing algorithms

In order to understand the process of feature selection of simulated annealing, first we start by using the basic of algorithm. In Table 2 shows the result of feature selection using both AIC and BIC

function and different temperature. We can see that for accuracy and specificity, both AIC and BIC have decreased but not significantly, contrary the sensitivity have been increasing but also relatively stable. This shown that by increasing the temperature, it would not significantly give different result. Compare between AIC and BIC function, while AIC function give more quite standard number on all goodness of the model, BIC function more high in accuracy and specificity but low in sensitivity which still not enough to capture the right variable. The sample result of this simulation can be seen in Figure 2.

Table 2. Results of basic simulated annealing algorithm

Initial Temperature	AIC			BIC		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
100	70.62%	60.80%	71.71%	87.74%	15.00%	95.82%
1000	67.39%	69.60%	67.14%	87.33%	14.50%	95.42%
10000	65.79%	72.60%	65.03%	86.66%	18.80%	94.20%
100000	61.85%	76.30%	60.24%	85.06%	22.50%	92.01%
1000000	60.83%	77.70%	58.96%	86.00%	18.80%	93.47%



3.2 Adjustment of simulated annealing algorithms

To increase the achievement of relatively good result from basic algorithm, it can be done by increasing the size of the neighbor. Table 3 shows the result of increasing the size of the neighborhood with temperature of 100000. In the table, we can see that the sensitivity of both functions reaches a perfect score of 100% which the good sign. However, in terms of accuracy and specificity from AIC function become more lower or getting worst. The different things happen to BIC function where even though the value of both is relatively the same as a value of more than 80% which also good prediction. The sample result of this simulation can be seen in Figure 3.

Because in AIC function the accuracy and specificity value are getting worst, then an improvement is made by adding up p-values to the process of selecting variables. It's mean that If all the variables are significant (p-value less than 0.05), then change variables are performed. Table 4 shows the result obtained with this method. We can see that either accuracy, sensitivity, and specificity values both AIC and BIC function reach 90% which good sign. The sample result of this simulation also can be seen in Figure 4

Table 3. Results of simulated annealing using increase neighborhood size

Neighborhood Size	AIC			BIC		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
5	56.64%	100.00%	51.82%	90.26%	100.00%	89.18%
10	46.40%	100.00%	40.44%	88.85%	100.00%	87.61%
20	38.83%	100.00%	32.03%	89.99%	100.00%	88.88%
50	32.11%	100.00%	24.57%	89.02%	100.00%	87.80%
100	25.99%	100.00%	17.77%	89.34%	100.00%	88.16%

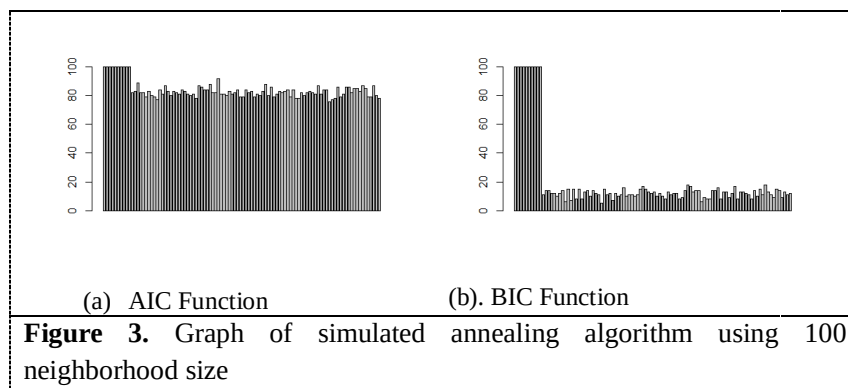
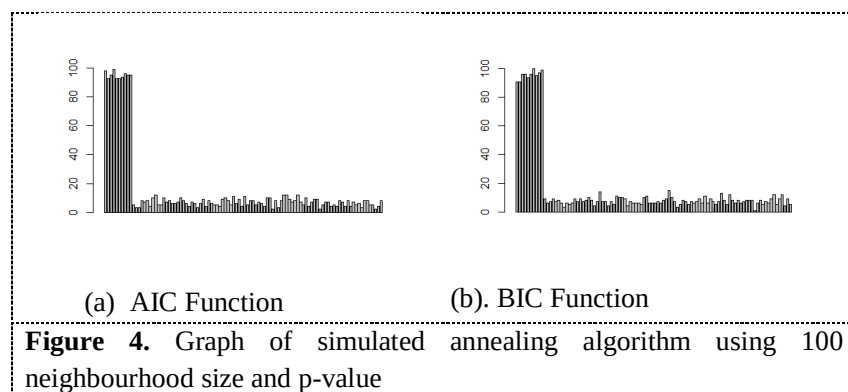


Table 4. Results of simulated annealing using increasing neighborhood size and p-value

Neighborhood Size	AIC			BIC		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
5	89.38%	61.20%	92.51%	90.39%	73.20%	92.30%
10	90.35%	67.10%	92.93%	90.36%	78.60%	91.67%
20	91.19%	73.90%	93.11%	90.47%	82.20%	91.39%
50	93.07%	86.80%	93.77%	92.38%	91.00%	92.53%
100	93.52%	95.10%	93.34%	92.92%	95.50%	92.63%



Even though using of p-value is good enough to select the variable, but it is still not optimal, a change in value is carried out by limiting the maximum of the selected variable to a value. This value indicates the number of variables that might be selected in the data. Table 5 shows the result obtained with this method which shows that the higher of neighbor value of both AIC and BIC functions, the accuracy and specificity value reaches 98% which means that this data show better result in predicting

the selected variables. The sensitivity also reaches 94% which also good in predicting the selected variables. The sample result of this simulation also can be seen in Figure 5.

Table 5. Results of simulated annealing using increasing neighborhood size and limited selection variable

Neighborhood Size	AIC			BIC		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
5	91.31%	56.50%	95.18%	96.64%	82.80%	98.18%
10	92.62%	63.00%	95.91%	97.60%	88.00%	98.67%
20	94.46%	72.30%	96.92%	98.02%	90.10%	98.90%
50	97.26%	86.30%	98.48%	98.40%	92.00%	99.11%
100	98.92%	94.60%	99.40%	99.18%	95.90%	99.54%

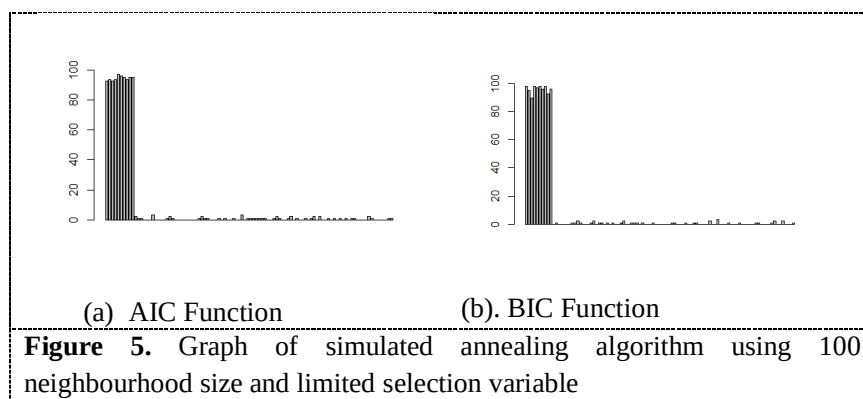


Table 6. Results of running time in simulated annealing algorithms using increasing neighborhood size, p-value, and limited selection variable

Neighborhood Size	AIC			BIC		
	Neighborhood Size	Neighborhood Size and P-Value	Neighborhood Size and Limited Selection Variables	Neighborhood Size	Neighborhood Size and P-Value	Neighborhood Size and Limited Selection Variables
5	5.28	3.14	3.28	5.21	5.16	2.81
10	9.83	5.89	6.18	7.95	9.12	6.40
20	23.04	11.36	13.36	16.88	20.45	13.16
50	44.43	28.10	29.96	38.19	40.39	32.04
100	60.80	56.90	60.29	60.17	60.21	60.87

The time needed for this method is also measured to see whether the process is faster or not compared to the general method. Table 6 shows the result of the three methods between using additional neighborhood size, neighborhood size and p-value, neighborhood size and limited selection variables in both AIC and BIC function which show that there are no significant differences between the three methods in both function. It also show that the improvement only occurs in the accuracy, sensitivity, and specificity in each method which is getting better.

4. Conclusion

Feature selection using simulated annealing can be used by adding neighborhood size and limited selection in algorithms using BIC function, but to estimate the number of limited selection can be seen by using BIC function with adding only neighborhood size.

Acknowledgments:

We are grateful to the Ministry of Research, Technology & Higher Education for the financial supported for this paper.

References

- [1] Sörensen K, Sevaux M, Glover F 2017 A History of Metaheuristics *arXiv:1704.00853v1*
- [2] Weisberg S 2005 *Applied Linear Regression: Third Edition* (New Jersey: John Wiley & Sons, Inc.)
- [3] Rutenbar R A 1989 Simulated Annealing Algorithm : An Overview. *Circuits and Device Magazine The Institute of Electrical and Electronics Engineers (IEEE)*, vol 5, p 19-26
- [4] Jeong I S, Kim H K, Kim T, Lee D H, Kim K J, Kang SH 2016 A Feature Selection Approach Based on Simulated Annealing for Detecting Various Denial of Service Attacks. *Convergence Security* vol 1 p 1-18
- [5] Akaike H 1974 A new look at the statistical model identification *IEEE Transactions on Automatic Control* vol 19 (6) p 716–723
- [6] Powers D M W 2011 Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation *Journal of Machine Learning Technologies* vol 2 (1) p 37–63