COMBINATION OF LOGISTIC REGRESSION AND SVM ALGORITHM WITH HYBRID PSO AND GA BASED SELECTION FEATURE IN CORONARY HEART DISEASE CLASSIFICATION

Sutrisno Situmorang[1],Pahala Sirait[2], Andri[3]
Master of Information Technology, STMIK Mikroskil, Medan, Sumatera Utara, Indonesia[1,2,3]
184212035@students.mikroskil.ac.id[1], [2]pahala@mikroskil.ac.id[2], [3]andri@mikroskil.ac.id[3]

abstract

Article Info
Received, 01 Mei 2021
Revised, 20 Mei 2021
Accepted, 20 June 2021

The world's high death rate from heart disease requires early prevention by medical doctors to diagnose heart disease early. The machine learning approach makes it possible to predict the risk of developing heart disease by examining certain values at a low cost. This study will contribute to the development of a combination of Logistic Regression and SVM models that integrate SVM and Logistic Regression algorithms by implementing selection features using hybrid PSO and GA methods. The combination concept of Logistic Regression SVM (LRSVM) applied to this study is to reduce the risk of SVM output errors by interpreting and modifying the output of SVM classifiers by the results of Logistic Regression analysis. The test results showed that LRSVM with pso-GA hybrid-based selection feature achieved better performance for coronary heart disease classification with 99.66% accuracy compared to classification accuracy with SVM algorithm without selection feature.

Keywords: SVM Logistic Regression Classification, Selection Feature, PSO-GA

1. Introduction

World Health Organization (WHO) data in 2012 showed 17.5 million people (31%) population dies from heart disease and that number will continue to increase. Meanwhile, data from *The Institute for Health Metrics and Evaluation* (IHME) in 2016 showed 17.7 million people (32.26%) people in the world died of heart disease. More than 3 million such deaths are under the age of 60 that are likely to be prevented early. A survey from the *Sample Registration System* in 2014 in Indonesia showed coronary heart disease was in the first place. For early prevention, some tests such as echocardiograms, nuclear *scans*, electrocardiograms (ECG), *angiography*, and stress tests are widely used by medical doctors to diagnose coronary heart disease. However, it requires expertise, risks for patients, costs are expensive and time-consuming (Kolukisa et al., 2019).

Bashir et al. (2019) does not specifically describe the specific algorithms used in feature selection (independent variables) nor does it specifically indicate the *Logistic Regression* (SVM) model mechanism as the best classification model. This research will contribute to the development of a *Logistic Regression* (SVM) model that integrates SVM and *Logistic Regression* algorithms by implementing selection features using hybrid PSO and GA methods. The concept of *Logistic Regression* (SVM) that will be applied to this study is to reduce the risk of SVM output errors by interpreting and modifying the output of SVM classifiers by the results of *logistic regression analysis.* That is, SVM class prediction depends on the result of logistic regression, i.e. if the result of logistic regression supports SVM output with high probability, then the SVM output does not change. Conversely, if the result of a logistic regression does not support SVM output with a low probability, the model changes the output of the SVM classifier.

Hybrid PSO and GA have been studied by Ali &Tauhid (2017) in the issue of optimization that minimizes the function of molecular energy. Their research studied the application of hybrid PSO and GA (HPSOGA) as *metaheuristic algorithms* due to the simplicity of PSO algorithms that are easy to implement, and require only a small number ofuser-defined parameters, but the drawbacks are inseparable from early convergence. GA is used to help PSOs regardless of early convergence. This study also contributed to the development of hybrid PSO and GA models by implementing 3 mechanisms. In the first mechanism, the PSO algorithm particle pool is applied with its strong performance with exploration and exploitation processes. The second mechanism is based on reducing the features and processes of population partitioning by dividing the population into sub-populations and implementing GA arithmetic operators i.e. crossovers on each subpopulation. Crossover implementation is expected to increase the search variation of the PSO algorithm. The third mechanism is to avoid early convergence by applying mutation operators of genetic algorithms throughout the population. The combination of these three mechanisms is expected to speed up the search and help the algorithm to achieve or approach the optimal solution in a reasonable time, where the results obtained help doctors to predict coronary heart disease with higher accuracy and make the right decisions.

2. Method

2.1 Data Mining

Data Mining is an implicit, previously unknown, and potentially useful extraction of information from data. The idea is to build a computer program that filters through the database automatically, looking for regularity or patterns. A strong pattern, if found, is likely to generalize to make accurate predictions on future data. Anything found will be imprecise: There will be exceptions for every rule and cases not covered by any rule. The algorithm must be powerful enough to cope with imperfect data and extract improper but useful regularity (Ian &Eibe 2005).

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build a computer program that sifts through the database automatically, looking for regularities or patterns. Strong patterns, if found, are likely to generalize to make accurate predictions on future data. Anything found will be inappropriate: There will be exceptions to every rule and cases are not covered by any rule. Algorithms must be robust enough to overcome imperfect data and extract imprecise but useful regularities (Ian &Eibe 2005).

2.2 Classification

Classification is one of the *Data Mining* techniques used to analyze a given dataset and retrieve each instance of it and assign this instance to a specific class so that the misclassification will be the least. This classification is used to extract models that accurately define important data classes in a given dataset. There are 2 stages in the classification process, the first step of the model is made by applying the classification algorithm to the training data set, and in the second step, the extracted model is tested against a predetermined test dataset to measure the performance and accuracy of the model trained by the model. So classification is the process of assigning class labels from datasets whose class labels are unknown (Nikam, *Orient. A. Comp. Sci. &Technol,*2015).

2.3 Logistics Regression

Logistic Regression is one of the machine learning techniques to classify records from datasets. *Logistic Regression* is an approach to create a prediction model as well as linear regression or commonly referred to as Ordinary Least *Squares* (OLS) *Regression*. The difference is that in logistic regression, researchers predict bound variables that scale dichotomy. The dichotomy scale in question is the nominal data scale with two categories, for example: Yes and No, Good and Bad, or High and Low (Virgeniya &Ramaraj, 2019).

2.4 Regression Analysis

Regression analysis is one of the analyses that aims to determine the influence of a variable on other variables. The simplest regression model is a simple linear regression model with the form of an equation:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

where:

Y = bound variable (predicted value)
X = free variable
β*0* = constant
β1 = regression coefficient (increase or decrease value)
ε = Random error.

2.5 Binary Logistics Regression Analysis

Logistic Regression is a statistical analysis method to describe the relationship between bound variables that have two or more categories with one or more category-scale or continuous free changes. The logistics regression can be divided into binary logistics regression, multinomial logistics regression, and ordinal logistics regression. The binary logistics regression model is used to analyze the relationship between one response variable and several predictor variables, with the response variable being a qualitative dichotomy data that is worth 1 to state the existence of a characteristic and worth 0 to express the absence of a characteristic. A binary logistics regression model is used if the response variable returns two categories of values 0 and 1, thus following Bernoulli's distribution as follows:

$$f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{2}$$

where:
πi = chance of the 1st occurrence
yi = i-i random modifier consisting of 0 and 1
The form of a logistic regression model with one predictor variable is:

$$\pi(x) = \frac{\exp\,(\beta_0+\beta_1 x)}{1+\exp\,(\beta_0+\beta_1 x)} \tag{3}$$

To make it easier to estimate the regression parameters, the π(*x*) in the equation above is transformed to produce a logit form of logit regression logistics, as follows:

$$g(x) = In\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \tag{4}$$

2.6 Genetic Algorithms

The genetics algorithm was first invented by JohnHolland, it can be seen in his book Adaption in Natural and Artificial Systems in the 1960s and then developed with his students and co-workers at the University of Michigan in the 1960s to 1970s. Holland's goal in developing Genetic Algorithms at that time was not to design an algorithm that could solve a problem, but rather to study the phenomenon of adaptation in nature and try to apply the mechanism of natural adaptation into computer systems. Genetic Algorithm as a branch of Evolution Algorithm is a method used to solve a value search in an optimization problem that is not linear problems (Savio &Chakraborty 2019).

2.7 Particle Swarm Optimization (PSO)

PSO is a *stochastic optimization* technique based on the social behavior of the movement of birds or fish and was first introduced by Russell C. Eberhart and James Kennedy in 1995. PSO has been successfully applied to various research fields as well as many applications, including specific applications with specific needs, such as function optimization, sudoku games, fuzzy system control, including "training" *Artificial Neural Networks* (ANN), solving supply chain problems (Habibi, 2017) After finding the two best values, update the particle speed and position with the following equation (Tuegeh et al., 2009):

$$v_{ij}^{k+1} = w_k + v_{ij}^k + c_1 * rand * (pbest_{ij}^k - x_{ij}^k) + c_2 * rand * (gbest_{ij}^k - x_{ij}^k) \tag{5}$$

$$v_{ij}^{k+1} = x_{ij}^k + v_{ij}^{k+1} \tag{6}$$

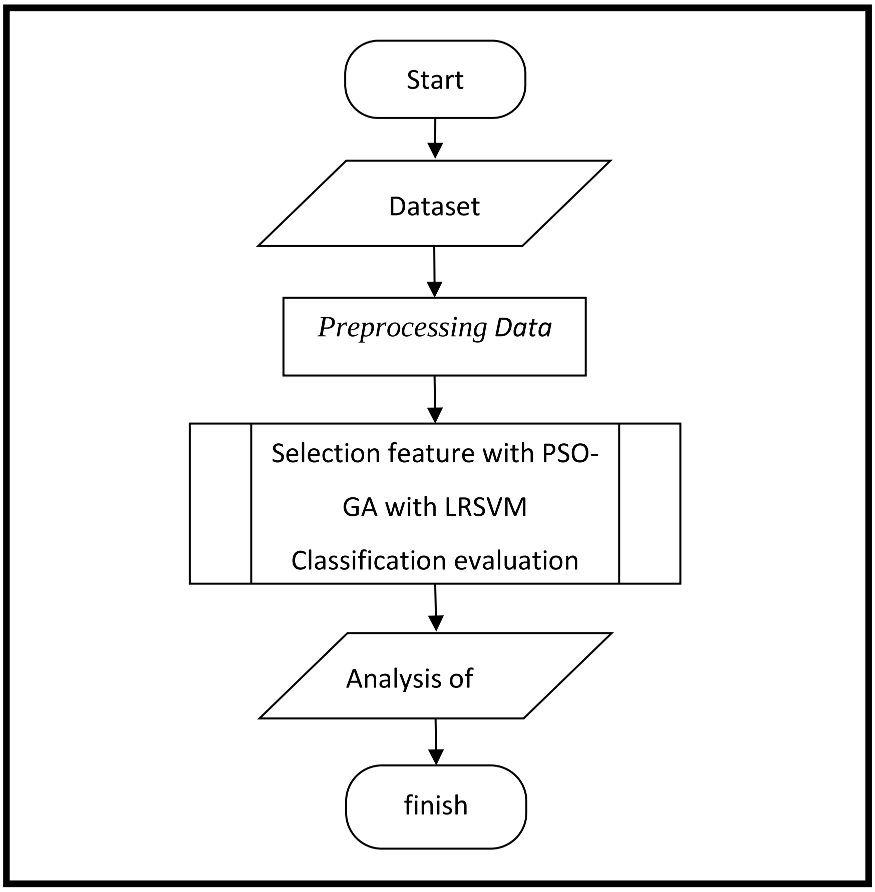

Figure 1. Flowchart Research Method

3. Results and Discussion

3.1 Preprocessing Data

The tests conducted in this study used datasets taken from the UCI *repository*, namely the Hungarian Institute of Cardiology data set, Budapest (Hungarian. data). A partial view of the Hungarian benchmark data set can be seen in Table 1.

Table 1. Table 1 Hungarian Benchmark Data Set View

No	attribute													class
	1	2	3	4	5	6	7	8	9	10	11	12	13	
1	28	1	2	130	132	0	2	185	0	0	2	0	6	0
2	29	1	2	120	243	0	0	160	0	0	2	0	6	0
3	29	1	2	140	251	0	0	170	0	0	2	0	6	0
4	30	0	1	170	237	0	1	170	0	0	2	0	6	0
5	31	0	2	100	219	0	1	150	0	0	2	0	6	0
6	32	0	2	105	198	0	0	165	0	0	2	0	6	0
7	32	1	2	110	225	0	0	184	0	0	2	0	6	0
8	32	1	2	125	254	0	0	155	0	0	2	0	6	0
9	33	1	3	120	298	0	0	185	0	0	2	0	6	0
10	34	0	2	130	161	0	0	190	0	0	2	0	6	0
11	34	1	2	150	214	0	1	168	0	0	2	0	6	0
12	34	1	2	98	220	0	0	150	0	0	2	0	6	0

No	attribute													class
	1	2	3	4	5	6	7	8	9	10	11	12	13	
13	35	0	1	120	160	0	1	185	0	0	2	0	6	0
189	44	1	4	130	290	0	0	100	1	2	2	0	6	1
190	46	1	1	140	272	1	0	175	0	2	2	0	6	1
191	47	0	3	135	248	1	0	170	0	0	2	0	6	1

Table 2. Best Accuracy Comparison of SVM and LRSVM Classifications

k-fold	**Akurasi Terbaik**	
	LRSVM PSO-GA	**SVM**
2-fold	99.66%	71.72%
5-fold	99.66%	77.59%
10-fold	99.66%	78.62%

Table 3. Best Accuracy Comparison of SVM and LRSVM Classifications

C	**F1 Score Terbaik**	
	LRSVM PSO-GA	**SVM**
0,1	98,69	67,40
0,5	99,30	67,23
1	99,12	71,50
5	99,01	66,07
10	98,77	67,14
50	98,81	64,48
100	98,40	64,48

4. Conclusions

This study developed a combination of Logistics Regression and SVM (LRSVM) classification techniques. The LRSVM concept applied using the technique lowers the risk of SVM output errors by interpreting and modifying the output of SVM classifiers according to the results of Logistic Regression analysis. Then a feature selection approach using the PSO-GA algorithm is applied to reduce the number of features. The performance of the proposed classification algorithm (LRSVM) is compared to the performance of conventional SVM classification algorithms without selection features by conducting tests on coronary heart disease benchmark data sets. The results showed that LRSVM with pso-GA hybrid-based selection feature achieved better performance with 99.66% accuracy for coronary heart disease classification compared to classification accuracy with SVM algorithm without selection feature by 78.62%, with an accuracy improvement of 21.03%. The LRSVM classification model with PSO-GA-based feature selection also achieved the highest F1 score of 99.30%, where a higher F1 Score indicates the performance of the proposed algorithm as an algorithm that performs better than the SVM classification model without the selection feature.

Reference

[1] Ali, A. F. & Tawhid, M. A., (2017). A Hybrid Particle Swarm Optimization And Genetic Algorithm With Population Partitioning For Large Scale Optimization Problems. *Ain Shams Engineering Journal*, Vol. 8, Issue 2, pp: 191–206.

[2] Eric R. Buhi, MPH, Ph.D.; Patricia Goodson, Ph.D.; Torsten B. Neilands, P.(2008). Out of Sight, Not Out of Mind: Strategies for Handling Missing Data. Handling Missing Data, 1(Handl. Missing Data), 83–92.

[3] Fariza, A., Martiana, E., Sucipto H. 2016. *Aplikasi Algoritma Genetika Multi Obyektif pada Traveling Salesman Problem*. Politeknik Elektronika Negeri Surabaya ITS.

[4] Gen, M. dan Cheng, R. 1997. Genetic Algorithms and Engineering Design.Newyork: Jhon Wiley & Sons, Inc.

[5] Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Canada: Addison-Wesley Publishing.

[6] Indira K & Kanmani, 2016. *Performance Analysis of Genetic Algorithm for Mining Association Rules*. International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012. Department of CSE, Pondicherry Engineering College Puducherry, 605014, India.

[7] Karegowda, A. G., Manjunath, A.S. & Jayaram, M.A., 2011. *Application of genetic algorithm Optimized neural network Connection weights for medical Diagnosis of Pima Indians diabetes*, International Journal on Soft Computing (IJSC), Vol.2, No.2, May 2011. Dept. of Master of Computer Applications, Siddaganga Institute of Technology, Tumkur, India.

[8] Kartheeswaran, S. & Durairaj, D. D. C. 2015. A Hybrid Genetic Algorithm and Back-Propagation Artificial Neural Network Based Simulation System for Medical Image Reconstruction in Noise-Added Magnetic Resonance Imaging Data. 2015 Online International Conference on Green Engineering and Technologies (IC-GET 2015).

[9] Liu, J. & Kong, X. 2018. Artificial Intelligence in the 21st Century. International Seminar Special Section On Human-Centered Smart Systems And Technologies IEEE 2018.

[10] Louridas, P. & Ebert, C. 2016. Machine Learning. Published By The IEEE Computer Society.

[11] Obitko, Marek. (1998). *Introduction To Genetic Algorithms*. University of Applied Sciences.

[12] Ray, S. 2019. A Quick Review of Machine Learning Algorithms. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019.

[13] Sivanandam, S. N. & Deepa S. N. 2015. *Introduction to Genetic Algorithms*. New York: Springer-Verlag Berlin Heidelberg.

[14] Sofge, D. A, 2002. Using Genetic Algorithm Based Variable Selection to Improve Neural Network Models for Real-World Systems. Proceedings of the 2002 International Conference on Machine Learning 7 Applications. Navy Center for Applied Research in Artificial Intelligence Naval Research Laboratory Washington, D.C., U.S.A.

[15] Savio, I & Chakraborty, U. K. 2019. Genetic Algorithm: An Approach on Optimization. Proceedings of the Fourth International Conference on Communication and Electronics Systems (ICCES 2019). https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124 dilihat tanggal 25 Oktober 2020
IEEE Conference Record # 45898; IEEE Xplore ISBN: 978-1-7281-1261-9

[16] Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A. & Bashir, K. 2019. Improving Heart Disease Prediction Using Feature Selection Approaches. Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan, 8th – 12th January 2019.

[17] Luxmi Verma1, L., Srivastava, S. & Negi, P. C., (2016). A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. *Journal Medical System*, Vol.40, Issue: 178, pp: 1-7.

[18] Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*(April)

[19] Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). A New Feature Selection Method To Improve The Document Clustering Using Particle Swarm Optimization Algorithm. *Journal of Computational Science*, Vol. 25, pp. 1-40.Fong, S., Wong, R., & Vasilakos, A. V. (2015). Accelerated PSO swarm search feature selection for data stream mining big data. *IEEE Transactions on Services Computing, Vol. 9(1).*

[20] Silva, P. H., Luz, E., Zanlorensi, L. A., & Menotti, D., (2018). Multimodal Feature Level Fusion

Based On Particle Swarm Optimization With Deep Transfer Learning. *IEEE Congress on Evolutionary Computation.*

[21] Xiao, Y., Wang, Y., & Sun, Y. (2018). Reactive Power Optimal Control Of A Wind Farm For Minimizing Collector System Losses. *Energies*.Tuegeh, M., Soeprijanto and Purnomo, M. H. (2009) 'Modified improved particle swarm optimization for optimal', *Seminar Nasional Aplikasi Teknologi Informassi 2009 (SNATI 2009)*, 2009(Snati), pp. 85–90.

[22] Virgeniya, S. C., & Ramaraj, E. (2019). Predictive analytics using rule-based classification and hybrid logistic regression(HLR) algorithm for decision making. *International Journal of Scientific and Technology Research*, *8*(10), 1509–1513.

[23] Kolukisa, B., Hacilar, H., Goy, G., Kus, M., Bakir-Gungor, B., Aral, A., Gungor, V. C., İşlem, K. B., Sistemleri, T., & Ankara, A. Ş. (2019). Diagnosis of Coronary Heart Disease via Classification Algorithms and a New Feature Selection Methodology. *International Journal of Data Mining Science*, *1*(1), 8–15.

[24] Tuegeh, M., Soeprijanto, & Purnomo, M. H. (2009). Modified improved particle swarm optimization for optimal. *Seminar Nasional Aplikasi Teknologi Informassi 2009 (SNATI 2009), 2009*(Snati), 85–90.