Credit Scoring Using CART Algorithm and Binary Particle Swarm Optimization

Reza Firsandaya Malik, Hermawan
Faculty of Computer Science,Universitas Sriwijaya, Indonesia

Article Info

Article history:

Received Aug 18, 2017
Revised Jun 18, 2018
Accepted Jul 15, 2018

Keyword:

CART
Credit Scoring
Data Mining
PSO

ABSTRACT

Credit scoring is a procedure that exists in every financial institution. A way to predict whether the debtor was qualified to be given the loan or not and has been a major concern in the overall steps of the loan process. Almost all banks and other financial institutions have their own credit scoring methods. Nowadays, data mining approach has been accepted to be one of the well-known methods. Certainly, accuracy was also a major issue in this approach. This research proposed a hybrid method using CART algorithm and Binary Particle Swarm Optimization. Performance indicators that are used in this research are classification accuracy, error rate, sensitivity, specificity, and precision. Experimental results based on the public dataset showed that the proposed method accuracy is 78 % and 87.53 %. In compare to several popular algorithms, such as neural network, logistic regression and support vector machine, the proposed method showed an outstanding performance.

Corresponding Author:

Hermawan,
Faculty of Computer Science,
Universitas Sriwijaya.
South Sumatera - Indonesia
Email: hermawan@mdp.ac.id

1. INTRODUCTION

Credit scoring is a particular job of loan lifecycle management that had been a big challenge. It predicts whether the debtor is qualified to be given a loan or not. Credit scoring is term that used to describe formal methods used for classifiying applicants for credit into good credit or bad credit classes. Indeed, wrong prediction will be a great loss to banks and financial institution. There are two types of misclassification pattern which is called type I and type II error [1]. Type I error occurs when the actually good credit, but later was not accepted and classified as bad credit which will reduce the institution’s profit. As the opposite, type II error occurs when the actually bad credit but later was classified as good credit. Thus, it will bring a big problem and serious damage to the institution [1]. With the increasing importance of credit scoring to bank and financial institution, this field has invoked interests to many researcher to work on it. This research area has been conducted by many researchers over years with so many methods. One of the very popular method is the data mining approach. Data mining has enticed a great importance of interest in the information industry in recent years that focused on the extraction of hidden knowledge from various data warehouse, data set, and data repositories [2]. This approach is a big help to bank and other financial institutions.

Some popular methods that had been used by some researcher are classification and regression tree (CART), Support Vector Machine (SVM), Artificial Neural Network (ANN), Multivariate Adaptive Regression Splines (MARS)[3]. Previously, researchers have used private dataset to explore credit scoring. For example, T. S. Lee, Chiu, Chou, and Lu have employed CART and multivariate adaptive regression splines (MARS) to private credit card local bank in Taipei, Taiwan. Experimental showed that compared to several algorithms, still CART and MARS have a better overall performance [4]. Another example, W. Chen et al have proposed hybrid method SVM + CART and SVM +multivariate adaptive regression splines (MARS) for their private dataset bank of China. Their results showed an improvement in term of accuracy using hybrid method [5]. Another researcher used public dataset for their experiment. J. Chen used german credit dataset and Australian from University of California (UCI) repository. He proposed a hybrid method called SVM + whitening space. His method showed an improvement compared to SVM[6]. Several approach using ensembles of classifier has been applied to credit scoring, such as bagging, boosting, random subspace, and decorate. The base classifier considered in the experimental study along with the ensembled methods are: logistic regression (LogR), multilayer perceptron (MLP), support vector machines (SVM), C4.5 decision tree (C4.5) and credal decision tree (CDT). From the result, credit decision tree as the base classifier has the better result, when it is used as base classifier, in a ensembled scheme for credit scoring assessment [7]. Almost all researcher works have focused their research on increasing the accuracy of credit scoring, such as Yao Ping, Lu Yongheng who proposed SVM + Neighborhood Rough Set and compared it with LDA, Logistic regression, Neural Network. Result shows that their proposed method gain an improvement in term of accuracy [8]. Some researchers, focused on catching “bad” creditors as an importance performance issue, with their proposed method Kernel Fuzzification Penalty - MCOC[9]. Other researchers, focused their work on time reduction for credit scoring, such as Bandhu & Kumar. Their work based on an approach called SVM + F Score sampling to reduced computational time for credit scoring and compared it with SVM + GA, Back Propagation and Genetic Programming.It is proved that their method is competitive, in the view of its accuracy as well as the proposed method has a less computational time[10]. Another issue is an imbalance datasets that became great concern by Hongliang He et al, that they focused their research on adaption of different imbalance ratios and proposed their novel method to obtain superior performance and high robustness[11].

In this paper, we proposed hybrid Classification and Regression Tree (CART) and Binary Particle Swarm Optimization. CART is well known specific decision tree algorithm. It is used in several kinds’ application of data mining, such as web mining, educational mining, medical mining, and credit scoring. Many researchers have employed CART in their investigation. One of their study using private dataset conclude that compare to some other popular intelligent methods such as SVM and Neural Network, CART shows a better performance in credit scoring in term of AUC measure[12]. CART has been admitted as one of top 10 data mining algorithm and one of the most influential data mining algorithm[13]. In contrast, Binary Particle Swarm Optimization (BPSO) as one of variant of PSO is used to increase overall performance of CART.

Particle Swarm Optimization is an algorithm, a kind of calculation method based on the theory of swarm intelligence, and a kind of model in the field of swarm intelligence that retains a global search strategy based on population of swarm[14]. With PSO, the problem is solved and addressed using swarm of particle that move around the swarm, looking for the best possible solution[15]. There are some advantages of using PSO such as, it does not need differentiation unlike many traditional method, and it has the ability to escape from local optimimum. Another advantages are PSO has flexibility to integrate with other optimization techniques in order to develop complex tools and it can be used for the objective functions with random nature, similar to the case that one of the optimization variables is random. Not to mention the fact tha that PSO has less sensitivity to the objective function’s nature, which means it has convexity or continuity [16]. Binary PSO is variant of Particle Swarm Optimization. It is a nature inspired algorithm, as well as metaheuristic global optimization algorithm, originally proposed by Kennedy and Eberhart. A type of bio-inspired optimization algorithm insipired by movement of birds and fish flock while searching for food [17]. PSO solution swam is compared to the bird swarm, the birds’ moving from one place to another is equal to the development of the solution swarm, good information is equal to the most optimist solution, and the food resource is equal to the most optimist solution during the whole course [18].This method has been used to several research area. It is used to classify high dimensional educational data with good performance result compare to several algorithms. Other researcher, embedded this method with SVM to analyze opinion mining of social media.Their study showed agood result, PSO affects the accuracy of SVM after the hybridization of SVM-PSO [19], [20]. Based on literature study, this method can also be used to improve overall performance of CART algorithm.

2. RESEARCH METHOD

Figure 1 shows flowchart of proposed research design. The following flowchart consists of sequence of steps and methods to do the research. It explains the process of conducting this experimental research in more details. Researchers will follow these steps while doing research to ensure the integrity of the whole research process.

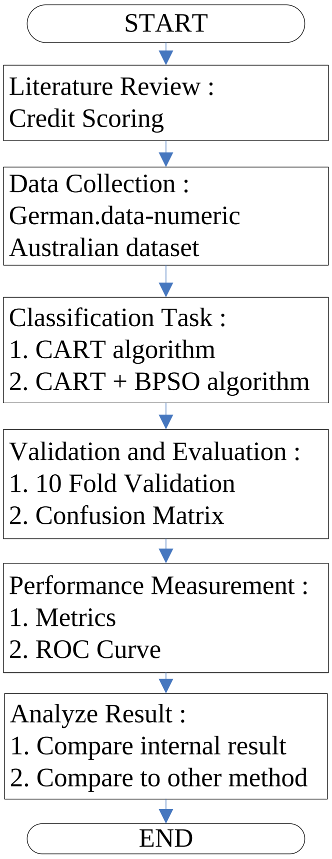

Figure 1. Proposed Research Design

Research began with collecting literature from few resources. A literature search conducted before proceeding to design experiment. This step provides foundational knowledge about the research area, the designs, instruments used, the procedure and the findings. The information discovered during this step helps the researchers fully understand the magnitude of problem. All materials were captured and extracted into research mapping. Later, we decided to use public dataset. Real world credit dataset, German.data-numeric dataset and Australian dataset are used as an object to our research. Considering the fact that based on our literature review, those datasets were generally used by researcher in the research area. The Datasets are available from the University of California (UCI) Repository of machine learning databases.The German.data-numeric dataset consists of 24 predictor attributes and 1 target attribute[21]. Total number of instances are 1000. There are 700 instances are labeled as creditworthy, and 300 instances are labeded as not creditworthy. Australian dataset consist of 14 predictor attributes and 1 target attribute.There are totally 690 instances in Australian dataset, consists of 307 instances are labeld creditworthy, and 383 instances are labeled as not creditworthy [21]. Table 1 further describes details of these datasets. The work of research is continued by conducted the classification task with CART algorithm and the proposed method (CART + BPSO). The experimental procedures will be carried out in this phase. Then 10-fold validation and confusion matrix are used to train our credit scoring model. Some metrics are used to measure performance of classifier. Metrics for evaluatingclassifier performance are accuracy, error rate, sensitivity, specificity, and precision.Overall performance is showed in Receiver Operating Characteristic (ROC) Curve and area under curve (AUC) of ROC[22], [23]. At last, our experimental result is analyzed and compared to the other similar method of data mining.

Table 1. Details of datasets

Dataset	No. attribut	No. Instances
German.data-numeric	25	1000
Australian	15	690

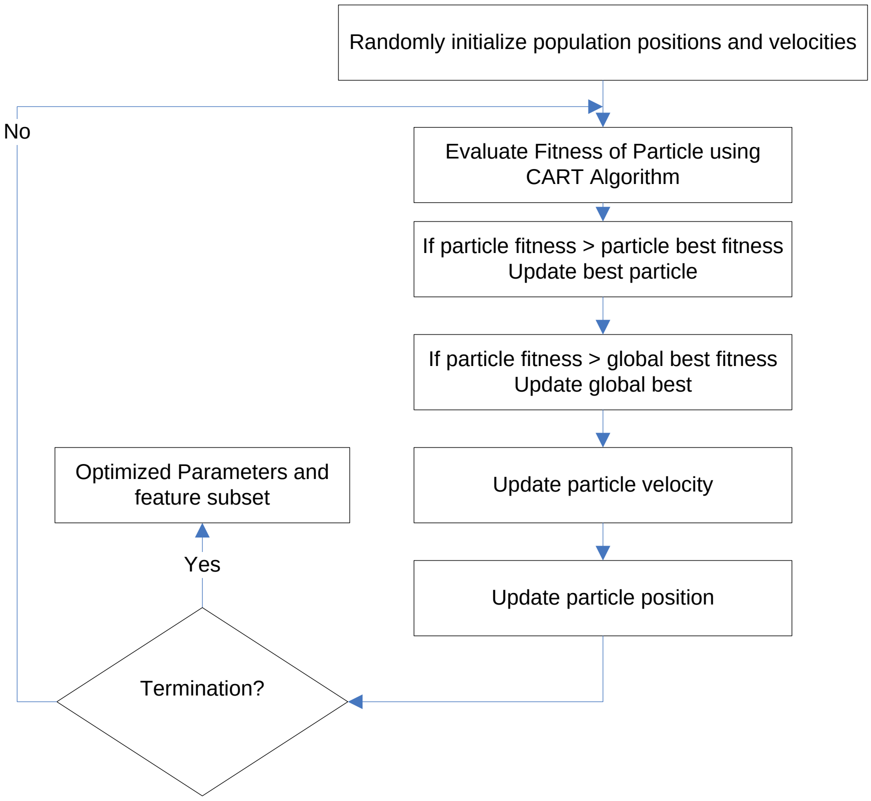
Randomly initialize population positions and velocities
Evaluate Fitness of Particle using CART Algorithm
If particle fitness > particle best fitness Update best particle
If particle fitness > global best fitness Update global best
Update particle velocity
Update particle position
Termination?
No
Yes
Optimized Parameters and feature subset

Figure 2. Proposed Method

Binary PSO approach is used as feature selection method to select best subset that produce best performance. BPSO is an extended algorithm of Particle Swarm Optimization that operates on binary search space. Each particle represents position in binary space and particle's position can take on the binary value 0 or 1. Figure 2 shows the flowchart of proposed method. It begins with randomly initialize particle. Population of particles are created, and each particle is correlated with generated solution. All particle's fitness is evaluated. This experimental study used CART classification accuracy as the fitness function. Based on the result, the next step is to evaluate particle's pbest and gbest. Followed by update particle velocity and sigmoid function. Construction phase let particles move to another potential solution based on its own experience and that of neighbor. The loop ended with a stopping criteria in termination phase that predetermined before[24][25].

3. RESULTS AND ANALYSIS

Experimental result is compared in two phase or part. First phase, an internal experimental result is compared each other. Performance of credit scoring using CART algorithm is compared to credit scoring using CART+PSO algorithm. Second phase, we compared proposed method to similar research. Table 2 shows the first phase comparison result.

Table 2. Comparison Result of CART and CART+BPSO

Metric	German.data-numeric dataset		Australian dataset	
	CART	CART+BPSO	CART	CART+BPSO
Accuracy (%)	75.2	78	85.36	87.53
Error rate (%)	24.8	22	14.64	12.47
Sensitivity (%)	89.1	91.71	84.04	86.97
Specificity (%)	42.7	46	86.42	87.99
Precision (%)	78.4	79.85	84.04	85.30
AUC	0.7196	0.7392	87.71	0.9034
No. of Attribute used	24	11	14	6

Table 3. Comparison result to other researchs

No.	Methods [German.data-numeric data set]	Accuracy %
	Support Vector Machine (SVM)	75.98
	SVM + Whitening Transformation (WT)	76.88
	Linear Disriminant Analysis	66.60
	Logistic Regression	72.40
	Neural Network	75.20
	SVM + Neighborhood Rough Set	76.60
	Multi-Criteria Optimization Classifier (MCOC)	73
	Kernel Fuzzification Penalty – MCOC	73.40
	SVM+ Genetic Algorithm	76.84
0	Back Propagation	76.69
1	Genetic Programming	77,26
2	Decorate + logR (ensemble)	77.40
3	Bagging + SVM (ensemble)	76.60
4	CART + BPSO (Proposed Method)	78

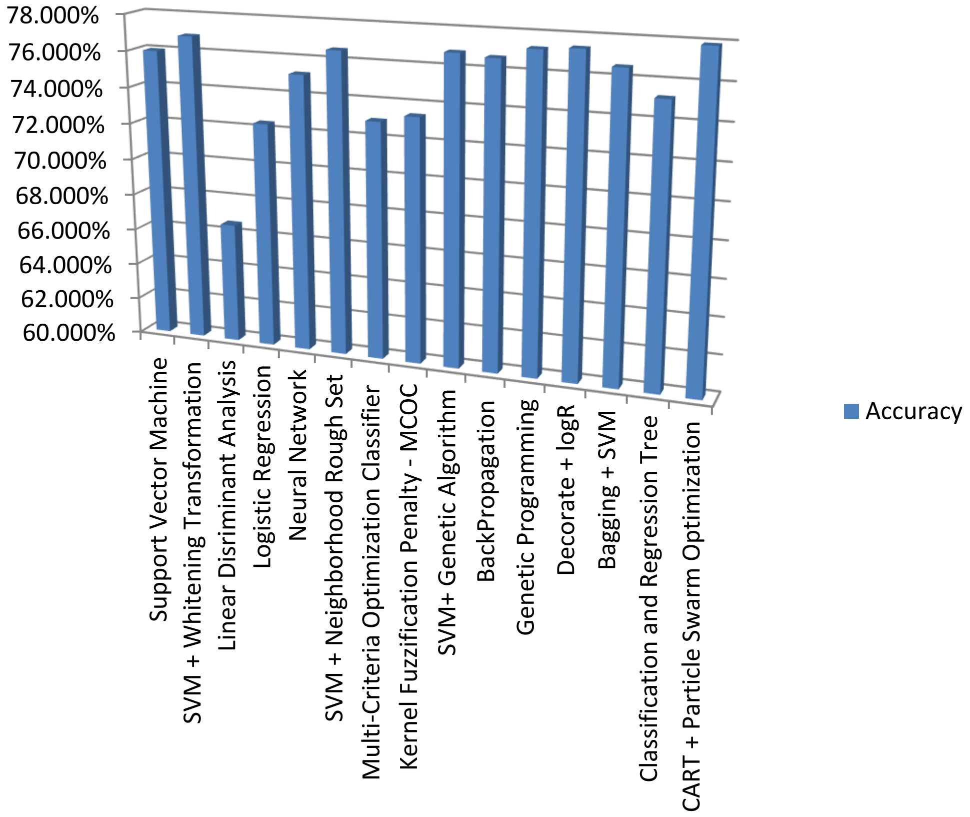

Figure 3. Accuracy comparison chart

Table 2 shows the overall performance of proposed method (BPSO+CART) compared to base method (CART). It is clear that there is remarkable improvement in the proposed method. Performance shows an increase in term of accuracy, the accuracy is raised from 75.2% to 78% with German.data-numeric dataset and 85.36% to 87.53% with Australian dataset. In term of error rate, proposed method shows a better performance. Another indicator of improvement, the area under curve (AUC) of our proposed method value is 0.7392 with German.data-numeric dataset and 0.9034 with Australian dataset, which are higher than the base learner method. Experimental result also shows that feature selection does affect overall performance. Feature selection is an importance task to improve the prediction accuracy of the hybrid model. Classification problems generally involve a number of features or attribute. However, not all of these features are equally important for classification task. Some of these features are not relevant and redudan. Our proposed method search for the most importance features from the search space (all features). CART + BPSO method used only 11 from 24 attributes and 6 from 14 attributes. The proposed method choose the best attribute that contribute the most to increase overall performance. Not to forget, the average execution time of our proposed method is about ten minutes. Term of execution or computational time is the next big challenge to our research, since speed has a great importance in the 21st century. The less computational time means more efficient and more benefit to the bank and industry.

Then we measured and compared our experiment result with another similar method and research. Figure 3 shows that compare to several well-known artificial intelligent and popular algorithm, our proposed method shows an outstanding result with 78 % accuracy. Accuracy level which is higher than Neural Network algorithm, Genetic algorithm and Support Vector Machine.

4. CONCLUSION

In this credit scoring research, we explore an approach to increase the performance of our base learner algorithm. CART algorithm is choosed as a base learner, since it is one of the best algorithms that is mostly used for the classification task. Binary Particle Swarm Optimization is adopted to increase the performance of CART algorithm. The proposed method is validated with real public credit dataset. The result shows an overall improvement of our experiment. Based on several indicators, the proposed method shows a better performance, such as accuracy, error rate, sensitivity, specificity and precision.

Compared to another research, our proposed method also shows an outperform result with 78 % accuracy, 22 % error rate with German.data-numeric dataset and 85.36 % accuracy, 14.64 % error rate with Australian dataset. Better classification rate than another popular classification algorithm such as support vector machine, neural network, and genetic algorithm. It also concluded the fact that feature selection as preprocessing step of data mining could increase performance.

Next big challenge is to increase the speed of execution of the proposed model, due to the long execution time. Since speed has become a problem, further research will be focusing to increase the speed of execution time. Future studies may use another feature selection method as part of fitness function BPSO.

ACKNOWLEDGEMENTS

This paper is part of research work for Master of Informatics, Faculty of Computer Science, Universitas Sriwijaya.

REFERENCES

[1] X.-L. Li, "An Overview of Personal Credit Scoring: Techniques and Future Work," *Int. J. Intell. Sci.*, vol. 02, no. 24, pp. 182–190, 2012.

[2] S. Hussain, N. A. Dahan, F. M. Ba-alwi, and N. Ribata, "Educational Data Mining and Analysis of Students ' Academic Performance Using WEKA," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 2, pp. 447–459, 2018.

[3] S. M. Sadatrasoul, M. Gholamian, M. Siami, and Z. Hajimohammadi, "Credit scoring in banks and financial institutions via data mining techniques: A literature review," *J. AI Data MiningJournal AI Data Min.*, vol. 1, no. 2, pp. 119–129, 2013.

[4] T. S. Lee, C. C. Chiu, Y. C. Chou, and C. J. Lu, "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines," *Comput. Stat. Data Anal.*, vol. 50, no. 4, pp. 1113–1130, 2006.

[5] W. Chen, C. Ma, and L. Ma, "Mining the customer credit using hybrid support vector machine technique," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7611–7616, 2009.

[6] J. Chen, "A Method of Improving Credit Evaluation with Support Vector Machines," *2015 11th Int. Conf. Nat. Comput.*, pp. 615–619, 2015.

[7] J. Abellán and J. G. Castellano, "A comparative study on base classifiers in ensemble methods for credit scoring," *Expert Syst. Appl.*, vol. 73, pp. 1–10, 2017.

[8] Y. Ping and L. Yongheng, "Neighborhood rough set and SVM based hybrid credit scoring classifier," *Expert Syst. Appl.*, vol. 38, no. 9, pp. 11300–11304, 2011.

[9] Z. Zhang, G. Gao, and Y. Shi, "Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors," *Eur. J. Oper. Res.*, vol. 237, no. 1, pp. 335–348, 2014.

[10] A. Bandhu and M. Kumar, "Computational time reduction for credit scoring : An integrated approach based on support vector machine and stratified sampling method," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 6774–6781, 2012.

[11] H. He, W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: adaption of different imbalance ratios Hongliang," *Expert Syst. Appl.*, 2018.

[12] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *J. Retail. Consum. Serv.*, vol. 27, pp. 11–23, 2015.

[13] X. Wu, V. Kumar, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, *Top 10 algorithms in data mining,* vol. 14, no. 1. 2008.

[14] J. He and H. Guo, “A Modified Particle Swarm Optimization Algorithm,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 10, pp. 6209–6215, 2013.

[15] W. A. Shukur and K. K. Jabbar, “Information Hiding using LSB Technique based on Developed PSO Algorithm,” *Int. J. Electr. Comput. Eng.*, vol. 8, no. 2, pp. 1156–1168, 2018.

[16] H. Shahinzadeh, S. M. Nasr-azadani, and N. Jannesari, “Applications of Particle Swarm Optimization Algorithm to Solving the Economic Load Dispatch of Units in Power Systems with Valve-Point Effects,” *Int. J. Electr. Comput. Eng.*, vol. 4, no. 6, pp. 858–867, 2014.

[17] R. F. Malik, T. A. Rahman, R. Ngah, S. Zaiton, and M. Hashim, “The New Multipoint Relays Selection in OLSR using Particle Swarm Optimization,” *TELKOMNIKA*, vol. 10, no. 2, pp. 343–352, 2012.

[18] Q. Bai, “Analysis of Particle Swarm Optimization Algorithm,” 2010.

[19] A. A. Yahya and A. Osman, “Classification of High Dimensional Educational Data using Particle Swarm Classification,” pp. 34–41, 2014.

[20] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, “Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization,” *Procedia Eng.*, vol. 53, pp. 453–462, 2013.

[21] M. Lichman, “{UCI} Machine Learning Repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml.

[22] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[23] F. Gorunescu, “Classification Performance Evaluation,” in *Data Mining. Concepts, Models and Techniques*, vol. 12, 2011, pp. 319–330.

[24] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, “Particle swarm optimization for parameter determination and feature selection of support vector machines,” *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1817–1824, 2008.

[25] S. Talukder, “Mathematical Modelling and Applications of Particle Swarm Optimization,” School of Engineering Blekinge Institute of Technology, 2011.

BIBLIOGRAPHY OF AUTHORS

Reza Firsandaya Malik was born in Padang, West Sumatera in 1976. He received his senior high school in SMAN 70 Bulungan, Jakarta (1991 - 1994). He graduated from Institut Sains dan Teknologi Nasional (ISTN), Jakarta, as S.T (Bachelor of Engineering) in 2000 and obtained M.T (Master of Technique) from Institut Teknologi Bandung in 2003. He received the PhD degree from Universiti Teknologi Malaysia (UTM) in 2011, where he investigated Routing Optimization Scheme in Wireless Mesh Networks using Particle Swarm Optimization.

He joined Faculty of Computer Science, Universitas Sriwijaya (UNSRI) as a Lecturer in December 2010. He also appointed as member of Communication Network and Security (COMNETS) Research Laboratory in Faculty of Computer Science, Universitas Sriwijaya. During completing Ph.D study in Wireless Communication Centre (WCC) (2004 - 2006), he involved in Wireless Campus Project – Design and Deployment of Hot-spot IEEE 802.11g Wireless LAN, collaboration between WCC, UTM and Industry. He worked closely as researcher in Malaysia government agencies such as Ministry of Science, Technology and Innovation (MOSTI) and Ministry of Higher Education (MOHE) Malaysia.

He appointed as a Co-Chief Editor in ComEngApp-Journal. Thus, as member of Institute of Electrical and Electronics Engineers (IEEE), mosharaka for research and studies (mosharaka.net) and Association of Informatics and Computer College (APTIKOM). His experience in journal management as a reviewer in TELKOMNIKA Journal, Journal of Network and Computer Applications (JNCA) and several International Conferences and also as Journal Editor in Computer and Engineering Applications (ComEngApp) and Institute of Advanced Engineering and Science (IAES). In UNSRI, his current research interests include computer networks and soft computing. He also assigned as Head of Service and Application Working Group in Indonesia 5G Forum.

Hermawanis a master student at the faculty of computer science University of Sriwijaya, Palembang, South Sumatera. Currently working as a lecture at information system department at STMIK GI MDP, Palembang, South Sumatera. Passionate about latest technology, developing information system, analyaze system. His research interest area in software engineering, data mining, data scientist, database and information system.