Improved Fuzzy K-Nearest Neighbor Using Modified Particle Swarm Optimization

Jamaluddin\textsuperscript{1}, Rimbun Siringoringo\textsuperscript{2}

\textsuperscript{1,2}Universitas Methodist Indonesia Jl. Hang Tuah No 8 Medan-Indonesia
\textsuperscript{1}jac.satuno@gmail.com

Abstract- Fuzzy k-Nearest Neighbor (FkNN) is one of the most powerful classification methods. The presence of fuzzy concepts in this method successfully improves its performance on almost all classification issues. The main drawback of FKNN is that it is difficult to determine the parameters. These parameters are the number of neighbors (k) and fuzzy strength (m). Both parameters are very sensitive. This makes it difficult to determine the values of ‘m’ and ‘k’, thus making FKNN difficult to control because no theories or guides can deduce how proper ‘m’ and ‘k’ should be. This study uses Modified Particle Swarm Optimization (MPSO) to determine the best value of ‘k’ and ‘m’. MPSO is focused on the Constriction Factor Method. Constriction Factor Method is an improvement of PSO in order to avoid local circumstances optima. The model proposed in this study was tested on the German Credit Dataset. The test of the data/The data test has been standardized by UCI Machine Learning Repository which is widely applied to classification problems. The application of MPSO to the determination of FKNN parameters is expected to increase the value of classification performance. Based on the experiments that have been done indicating that the model offered in this research results in a better classification performance compared to the Fk-NN model only. The model offered in this study has an accuracy rate of 81%, while. With using Fk-NN model, it has the accuracy of 70%. At the end is done comparison of research model superiority with 2 other classification models; such as Naive Bayes and Decision Tree. This research model has a better performance level, where Naive Bayes has accuracy 75%, and the decision tree model has 70%.

Keywords: fuzzy k-nearest neighbor, modified particle swarm optimization, german credit data

1. Introduction

Non-performing loans is one of the most important issues on industrial and financial services \cite{1}. At a certain level, accumulation of loan defaults can trigger bankruptcy of a bank and other financial institutions. Learn about the background and characteristic of financial customers a factor that significant before deciding to give credit. To answer these problems, one concern of funding or banks is how to build an assessment technique to ensure eligibility of a customer on the filing of credit \cite{2}. Machine Learning is a field of information technology that has been widely applied to build decision support system, especially in the field of economics and finance. Machine learning plays a very important role and has resulted in various research studies that have been applied to related to credit risk assessment. The following is a variety of machine learning-based research that has succeeded in building the instruments and models used in the effective credit risk predictors and estimators. The researches are: K-Nearest Neighbor Method \cite{3}, \cite{4}, Fuzzy K-Nearest Neighbor \cite{5}, Bayesian
Network [6], Support Vector Machine [7], Artificial Neural Network [8], Fuzzy Immune Learning [9] and Logistic Regression [10].

The k-Nearest Neighbor (k-NN) method is the most popular machine-learning method, simple and easy to implement (Wang & Li, 2010), k-NN has two weaknesses. Firstly, the success of this method depends on the number of neighbors applied, so in order to produce a high degree of accuracy, researchers should try different values of k with varying amounts, of course this is not effective because it is done manually. This may be reflected in the research conducted [3], by applying the various k, the best accuracy is obtained at k = 3, while [5] obtains the best accuracy at k = 13. Second, in addition to the reliance on the k value, the relation between data with classes is rigid (crisp), where each instance/event/incident has only a relationship with one class exclusively, whereas with the other classes, it has no relationship at all.

Many attempts have been made to avoid the k-NN stiffness properties. One attempt is to combine fuzzy theory into k-NN. The application of the fuzzy theory produced a new method known as Fuzzy k-Nearest Neighbor (Fk-NN) [11]. In the Fk-NN model, the relationship between data and classes is not crisp, each data and class has membership relationship with a certain level. The strength of the relationship requires a fuzzy strength (m) parameter. Compared to k-NN, Fk-NN always results in a higher degree of accuracy in all classification problems [12]. It is also the main reason why the Fk-NN method becomes the preferred method of interest in many researches, especially related to the economic issues. The fuzzy strength parameters (m) and the neighboring number parameters (k) are fundamental determinants of the Fk-NN model. It means that the values of m and k have a direct impact on the model accuracy. Determining the values of ‘m’ and ‘k’ is often not easy and difficult to control because there is no theory or guide that concludes how appropriate ‘m’ and ‘k’ should be [12]. To answer the problem, it is necessary to have another method that can help the Fk-NN model to find the value of m and k.

In this study, we offer an approach to assess the optimization parameter solution in order to determine the best value of m and k. In this study, we applied Modified Particle Swarm Optimization (MPSO). The application of the method is based on some considerations. The first consideration, compared with similar algorithms, such as Genetic Algorithm (GA), PSO is relatively simple because it does not have many procedures such as selection, mutation and crossover procedures in GA. The second one, the PSO method has been proven to optimize the parameters of other machine learning methods. This is shown through the following studies. PSO is well-suited in combination with SVM [13], PSO and Artificial Neural Network [14], PSO with Self Organizing Map or SOM [15]. The results obtained in these models indicate improved accuracy through the application of the PSO in optimizing parameters. In this study, MPSO (Modified Particle Swarm Optimization) which is another variant of PSO was used to optimize Fk-NN parameters. This research built a model to evaluate the granting of credit based on FK-NN and MPSO classification. In other words, optimization of Fk-NN parameters by MPSO is expected to increase classification accuracy.

2. Research model

The FKNN-MPSO Classification Model This research was conducted in several stages, including pre-process data stage, selection of parameters k, m optimal using MPSO and FKNN. The procedures can be explained in the following steps:

Step 1 : Initialize the values of ‘k’ and ‘m’. In the FK-NN model, the values of ‘k’ and ‘m’ must be firstly initialized.
Step 2: Generate the initial particles randomly. The values of ‘k’ and ‘m’ are generated in parallel as many as 8 particles.

Step 3: The k and m values above are used to train the data.

Step 4: Each particle is evaluated based on the fitness value. The fitness function is used to calculate fitness or the level of goodness of an individual to survive. This function takes individual parameters and produces the fitness value of the individual. For each problem to be solved with the PSO algorithm must be defined by a fitness function. In this study, the fitness function is presented in the equation (8) = ( + ) ( + + + ) (8)

Step 5: Update the position and the velocity of the particles. The PSO procedure requires that the position and the velocity of each particle be evaluated by equations (2.11) and (2.12).

Step 6: At this stage, again testing FK-NN model to find the best particle. The best particles evaluated by the fitness value in step 4 were applied to train the training data and to calculate its fitness value.

Step 7: Update personal optimal fitness (pfit) and personal optimal position (pbest). So far all the particles in the first iteration have been evaluated, the first iteration yields the best particle (pbest).

Step 8: At this stage it is evaluated whether all iterations have been completed. If it is not completed yet, resumed in the step 3, if it is completed then proceed to step 10.

Step 9: Each iteration has the best particles. At this stage the best particles of all iterations will be evaluated to determine the best global particles optimal position (gbest).

Step 10: If the gbest value in step 10 is the expected value then at that moment has obtained the optimum k and m value and continued to step 12. If the gbest value has not matched the expected criterion yet, then the new resurrected population is back to the step 4.

Step 11: Perform data classification testing by entering the data test/the tested data.

Step 13: Cross validation to see if the classification result in the step 12 has the best accuracy. If yes then the whole process is completed, if not then do the evaluation the value of ‘k’ until the smallest accuracy value is obtained.

3. Experimental Design

A. Dataset

The dataset used in this study is German Credit Data, from UCI Machine Learning Repository. This dataset was chosen for several considerations. Firstly, it does not contain missing values [22]; secondly, this dataset also belongs to a high-dimensional dataset with 21 attributes (A1 through A21) and 1000 data. The dataset consists of 1000 data, 21 attributes including class attributes. Datasets are grouped into 2 classes, i.e. 700 instances for the good classes and 300 instances for the bad ones. This dataset does not contain missing value or an empty value [23]. Table 2 shows the description of the dataset and the description of its attributes is shown. The dataset are available in the .arff format (attribute relation file format attribute).

| Number | Dataset          | German Credit Data |
|--------|------------------|--------------------|
| 1      | Attribute type   | Category           |
| 2      | Number of attribute | 20                |
| 3      | Number of instance | 100               |
| 4      | Missing value    | No                 |
| 5      | Number of class  | 2                  |
Table II German credit data description

| Dataset | German Credit Data | Attribute type | Category and number of attributes | Number of instance | 1000 | Missing value | No | Number of class | 2 |

B. Modified Particle Swarm Optimization

1) Parameter Setting: The MPSO (Modified Particle Swarm Optimization) algorithm is a parametric method, where its application requires parameter setting. In Table III below is given a list of parameters of MPSO applied.

2) Particle Encoding: A particle encoding scheme is performed by generating random numbers ranging from 0 to 10 with a length of 20 numbers. Of the 20 numbers, the numbers 1 through 15 are used for the generation of parameters m, numbers 16 to 20 are applied for the generating parameters k. Figure 2 shows a particle coding scheme.

![Particle coding scheme](image)

Figure 1. Particle coding scheme

4. Research Results

MPSO algorithm begins with swarm initialization process. In the table IV, the swarm initialization scheme is displayed. The table segmentation consists of 3 parts: particle composition, parameters (k, m) and fitness value. Swarm consists of 10 particles (P1-P10), each particle contains of 20 random numbers, k and m values and the fitness. For example, particle 1 consists of 20 random numbers (7 1 2 1 3 5 3 9 0 4 4 6 9 2 4 0 9 7 0). The value of k, m which is raised on the particle is 13, 2.31, and the value of k, m obtained the fitness value 0.76 or 76%. Update the particles: In the next swarm, each particle will be updated until it reaches the target fitness value of 1.00 (100%). If this value can not be achieved then the highest value will be taken as a solution. Table V displays the particle to swarm-100, and resulted in the improvement of k, m and fitness values. Renewal of k and m value and fitness value improvement Based on the Table V below, the best fitness value is 0.81 (81%) and the value of k, m (31, 13.23).

![Table 2](image)

Table 2. Global Best

| S  | GBest | k, m | fitness |
|----|-------|------|---------|
| 100| 91    | 31, 13.23 | 0.81    |

Global Best (GBest): Global Best or GBest is the particles with the greatest fitness value of all swarm available. Table VI shows GBest. The table contains information about swarm, particle and fitness values (Global Best fitness). The GBest position is achieved in the 100th swarm (S100), k and m values (k, m) are 31, 13.23 and the fitness is 0.81 (81%). Table III Global Best S GBest k, m fitness. 100 91 85 86 85 87 89 90 89 90 93 86 88 84 93 91 84 31, 13.23 0.81
Influence of the Particle Number: To examine whether there is an effect of applying the number of particles to the increase of fitness value (in this case is the accuracy), this study applied the number of particle varies i.e 10 particles, 20 particles and 40 particles. The results are shown in the table VII, below. Based on the three tests, it can be indicated that the increase in the number of particles has no significant effect on fitness improvement.

Table 3. Test Result 10 Particle 20 Particle 40 Partikel No. k,m fitness No. k,m fitness No. k,m fitness

| No  | 10 Particle |  | 20 Particle |  | 40 Particle |  |
|-----|-------------|---|-------------|---|-------------|---|
| 1   | 15, 3.09    | 0.76 | 9, 3.87    | 0.77 | 13, 3.87    | 0.78 |
| 2   | 15, 3.09    | 0.76 | 17, 4.65   | 0.77 | 13, 3.87    | 0.78 |
| 3   | 15, 5.43    | 0.79 | 17, 6.21   | 0.77 | 19, 3.87    | 0.78 |
| 4   | 31, 12.45   | 0.84 | 23, 9.33   | 0.77 | 25, 6.21    | 0.84 |
| 5   | 15, 7.77    | 0.84 | 25, 9.33   | 0.77 | 29, 10.89   | 0.84 |
| 6   | 17, 7.77    | 0.84 | 25, 10.11  | 0.84 | 29, 10.89   | 0.84 |
| 7   | 23, 8.55    | 0.84 | 31, 10.11  | 0.84 | 31, 12.45   | 0.84 |
| 8   | 27, 9.33    | 0.84 | 31, 10.11  | 0.84 | 31, 12.45   | 0.84 |
| 9   | 31, 11.67   | 0.84 | 31, 10.11  | 0.84 | 31, 12.45   | 0.84 |
| 10  | 31, 11.67   | 0.84 | 31, 10.11  | 0.84 | 31, 12.45   | 0.84 |
| 11  | 31, 12.45   | 0.84 | 31, 10.11  | 0.84 | 31, 12.45   | 0.84 |
| 12  | 31, 13.23   | 0.84 | 31, 10.11  | 0.84 | 31, 13.23   | 0.84 |
| 13  | 31, 13.23   | 0.84 | 31, 13.23  | 0.84 | 31, 13.23   | 0.84 |
| 14  | 31, 13.23   | 0.84 | 31, 13.23  | 0.84 | 31, 13.23   | 0.84 |
| ... | ...         | ... | ...        | ... | ...         | ... |
| 100 | 31, 13.23   | 0.84 | 31, 13.23  | 0.84 | 31, 13.23   | 0.84 |

6) Global Seeking: Global Seeking is the GBest tracing process starting from the first swarm to the last swarm. In this penance GBest is the fitness with the greatest value, so the search process will form the ascending graph as shown in figure 3.

C. Testing Performance model classification Fk-NN + MPSO

The results showed that the Fk-NN + MPSO model resulted in better classification performance than the Fk-NN model. The comparison of both performance model is presented in the table X and Figure 4.

Table 4. Comparison of Fk-NN model performed with Fk-NN + MPSO model

| Performance (0 to 1) | Model | Fk-NN | Fk-NN + MPSO (This Experiment) |
|---------------------|-------|-------|--------------------------------|
| Accuracy            | 0.71  | 0.81  |
| Precision           | 0.78  | 0.93  |
| AUC                 | 0.77  | 0.84  |

Fk-NN + MPSO (This Experiment) Accuracy 0.71 0.81 Precision 0.78 0.93 AUC 0.77 0.84
D. The Comparison with Superiority Models

To validate the superiority of the FK-NN + MPSO model in predicting creditworthiness, the performance of FK-NN + MPSO method is compared to other models; the Naive Bayes method and the Decision Tree method. The performance results using the Naive Bayes method, the Decision Tree method is presented in Table XI. Figure 5 shows the comparison in a graphical form.

Table 5. Comparison of model superiority

| Model       | Fk-NN | Naïve Bayes | Decision Tree | Fk-NN + MPSO (This Experiment) |
|-------------|-------|-------------|---------------|---------------------------------|
| Accuracy    | 0.71  | 0.75        | 0.70          | 0.81                            |
| Precision   | 0.78  | 0.74        | 0.68          | 0.93                            |
| AUC         | 0.77  | 0.78        | 0.64          | 0.84                            |

5. CONCLUSION

This research offers a model of Fuzzy k-Nearest Neighbor (Fk-NN) classification based on the Modified Particle Swarm Optimization (MPSO) algorithm. The application of MPSO to Fk-NN aims to improve classification performance through the optimization of adjacent (k) and fuzzy (m) parameters. MPSO application can eliminate the aspect of subjectivity in determining parameters k and m. The result of the research shows that the MPSO application in the classification process has improved the accuracy of Fk-NN, thereby affecting the performance improvement of creditworthiness classification. The Fk-NN and MPSO models can work on high-dimensional data such as germancredit datasets. This study also compared the superiority of the FK-NN + MPSO model with other classification models namely Naive Bayes and Decision Tree. The result shows that FK-NN + MPSO model provides superior performance compared to the two models.

6. BIBLIOGRAPHY

[1] Lee, M-C. Enterprise Credit Risk Evaluation models: A Review of Current Research Trend. International Journal of Computer Applications, 44(11) : 0975 – 8887.2012
[2] Ghatasheh, A. Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study. International Journal of Advanced Science and Technology, 72 (2014) : 1930.2014.
[3] Abdelmoula, A. K. Bank credit risk analysis with k-nearest neighbor classifier: Case of Tunisian banks. Accounting and Management Information Systems, 14 (1) : 79-106.2015.
[4] Rahman, M.M., Ahmed, S. & Shuvo, M.H. Nearest Neighbor Classifier Method for Making Loan Decision in Commercial Bank. I.J. Intelligent Systems and Applications, 4 (8) : 60-68.2014
[5] Kurama, O., Luukka, P. & Collan, K. Credit Analysis Using a Combination of Fuzzy Robust PCA and a Classification Algorithm. Advances in Intelligent Systems and Computing-Springer, 3(15) : 19-29. 2015.
[6] Mortezapour, R. & Afzali, M. Assessment of Customer Credit through Combined Clustering of Artificial Neural Networks, Genetics Algorithm and Bayesian Probabilities. International Journal of Computer Science and Information Security, 11(12) : 1-5. 2013.
[7] Chen, Q., Xue, H.F. & Yan, L. Credit risk assessment based on potential support vector machine. International Conference on Natural Computation (ICNC) : pp. 1-25. 2011.

[8] Karimi, A. Credit Risk Modeling for Commercial Banks. International Journal of Academic Research in Accounting, Finance and Management Sciences, 4 (3) : 1-6. 2014.

[9] Kamalloo, E. & Abadeh, M.S. Credit Risk Prediction Using Fuzzy Immune Learning. Advances in Fuzzy Systems-Hindawi, 3 (2014) : 1-12. 2014

[10] Takyar, S.M.T., Nashtaei, R.A. & Chirani, E. The comparison of credit Risk between Artificial Neural Network and Logistic regression Models in Tose-Taavon Bank in Guilan. International Journal of Applied Operational Research, 5(1) : 63-72. 2014.

[11] Rosyid, H., Prasetyo, E. & Agustin, S. Perbaikan Akurasi Fuzzy K-Nearest Neighbor In Every Class Menggunakan Fungsi Kernel. Prosiding Seminar Nasional Teknologi Informasi dan Multimedia 2013 : pp. 13-18. 2013.

[12] Derrac, J., Chiclana, F., Garcia, F. & Hererra, F. Evolutionary Fuzzy K-Nearest Neighbors Algorithm using Interval-Valued Fuzzy Sets. Centre for computation intelligent : 1-28. 2014.

[13] Danenas, P. & Garsva, G. Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach. Proceeding of International Conference on Computational Science-ICCS : pp. 1324 – 1333. 2012.

[14] Li, S., Zhu, Y., Xu, C. & Zhou, Z. Study of Personal Credit Evaluation Method Based on PSORBF Neural Network Mode. American Journal of Industrial and Business Management, 3(2013) : 429-434. 2013.

[15] O’Neill, M. & Brabazon, A. Self-Organizing Swarm (SOSwarm) for Financial Credit-Risk Assessment. Proceeding on 2008 IEEE Congress on Evolutionary Computation : pp. 3087 – 3093. 2008.

[16] Keller, M. James, Michael R Gray, James A. Givens. A Fuzzy K-Nearest Neighbor. IEEE Transactions On Sistem and Cybernetics, 15(4) : 1-8. 1985.

[17] Jacubcoca, M., Maca, P. & Pech, P. A comparison of selected modifications of the particle swarm optimization algorithm. Journal of Applied Mathematics., 14(2014) : 10-15. 2014.

[18] Guo, H. & He, J. A modified particle swarm optimization algorithm. Journal of Computer Science10(2) : 341-346.2013.

[19] Yang, C.H., Hsiao, C-H. & Chuang, L-Y. Linearly decreasing weight particle swarm optimization with accelerated strategy for data clustering.International Journal of Computer Science, 37(3) : 3-9. 2010

[20] Clerc, M. The swarm and the queen: towards a deterministic and adaptive particle swarm optimization. Proceeding of Congress on Evolutionary Computation : pp. 19511957.1999.

[21] Han, J., Kamber, M. & Pei, J. Data mining tecniques and concepts. Morgan Kaufman publisher. Watham : USA. 2012

[22] Ramya, R., S. Analysis of feature selection techniques in credit risk assessment. Proceedings of International Conference on Advanced Computing and Communication Systems (ICACCS -2015) : pp. 1-6. 2015