Jurnal Mantik
Volume 4, Number 2, Aug 2020, pp. 1417-1423 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Optimization Performance of Fuzzy K-Nn with Modifield Particle Swarm Optimization in Credit Risk Classification

[1]Wita Clarisa Ginting, [2]Ronsen Purba, [3]Arwin

[1,2,3]Department Teknologi Informasi,
STMIK Mikroskil, Jl. M.H Thamrin No.140 Kel, Pusat Ps., Kec. Medan Kota, Kota Medan, Sumatera Utara 20212, Indonesia.

e-mail: [1]witaclarisa.ginting@gmail.com, [2]ronsen@mikroskil.ac.id, [3]arwin@mikroskil.ac.id

ARTICLE INFO

Article history:
Received: 12/01/2020
Revised: 22/07/2020
Accepted: 01/08/2020

Keywords:
Classification, Fuzzy K-NN, Particle Swarm Optimization, Credit Risk.

ABSTRACT

Credit risk is a risk due to the failure or inability of the customer to return the amount of credit obtained from the company and its interest according to a predetermined or scheduled period of time. The main task of the credit risk classification method is to provide a separation between those who have the potential to fail and those who have not failed in terms of credit payments. The k-Nearest Neighbor (kNN) method as the most popular, simple and easily implemented machine learning method can be used to classify credit risk. However, its success depends on the number of neighbors or neighbors (k) applied and the relationship between each data with a class is rigid (crisp) where each data only has a relationship with one class exclusively, while the other classes have no relationship at all. This study proposes the incorporation of the principles of fuzzy logic into k-NN to minimize the stiffness that results in a new method known as Fuzzy k-Nearest Neighbor or Fk-NN. However, the fuzzy strength factor (m) and the number of neighbors (k) as the fundamental determinants of Fk-NN which have a direct impact on the accuracy generated by the model, the determination is often not easy and difficult to control, so the Modified method is proposed Particle Swarm Optimization (MPSO) to be able to help Fk-NN find the best m and k values non-manually. The results of the classification of credit risk data are 1000 data, with 900 composition of training data (90%) and 100 data (10%) of test data using Fk-NN with MPSO producing accuracy reaching 92.4%, with the best k value is 7 and the best m value is 9.

1. Introduction

Credit crunch describes a situation where the agreement refund of the credit risk of failure, and shows that the company will be obtaining the loss potential (Leidiyana, 2013).

Credit risk also called default risk is a risk due to the failure or inability of the customer to return the amount of credit obtained from the company plus interest in accordance with a predetermined time period or scheduled (Tri, 2011).

The method of classification of the level of credit risk an important contribution to the key loan approval process that accurately and efficiently measure the level of credit risk of prospective borrowers. The method of classification of credit aims to predict the behavior of future in terms of credit risk based on the experience of past customers with similar characteristics. The level of credit risk of borrowers associated with a possible default risk on loans that it approved at a predetermined time (Lopez & Jeronimo, 2015).

The main task of classification methods of credit risk is to provide separation between them that could potentially fail with that did not fail in the case of payment of the credit. The ability of separation is the main indicator of the success of a method (Abellán& Castellano, 2017; Ala &Abbod, 2016).

The method of k-Nearest Neighbor (kNN) is a method of Machine Learning which is the most popular, simple and easy diimpelentasikan. In addition to these advantages, kNN has two drawbacks.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

First, the success of this method depends on the number of neighbors or neighbor (k) is applied, so as to produce a high level of accuracy, should be tried as the value of k with the number of which varies, of course, it is not effective because it is done manually. This can be reflected in the research conducted Moula (2005), by applying k vary, the level of the best accuracy is obtained at k=3, while Kurama et al. (2015) obtain the best accuracy at k=13. Second, in addition to the dependence on the value of k, the relationship between each data class is rigid (crisp) where each data only has a relationship with one class exclusively, while the other class does not have a relationship at all.

Various attempts have been made to avoid properties the stiffness of the k-NN. One of the efforts to minimize such rigidity is by combining the principles of fuzzy logic into k-NN. The merger resulted in a new method known as Fuzzy k-Nearest Neighbor or Fk-NN (Rosyid et al, 2013). In the Fk-NN the relationship between the data classes are not rigid, each class and data have a relationship of membership or membership with a certain degree. The strength of such a relationship requires the parameters of the fuzzy strengh (m). Compared with k-NN, Fk-NN generates the achievement of a higher degree of accuracy in almost all the problems of classification (Takyar et al, 2014).

Factors kekatan fuzzy (m) and the factor of Jumba's neighbors (k) is the factor determining the stature on the fundamental group of FK-NN, artinyaimsdampasung rice pakiakursung get along asehasio d replace the model. Penent the value of m and K often is not easy and is difficult to control because there is a theory or a menyulultberharusnya the value of M right (Rosyid et al, 2013).

To answer the above problems, it is necessary to include other methods which can help the Fk-NN find the value of m is. In this study the author offers an approach to the solution of parameter optimization (parameter optimization) in order to award the value of m and k adaptive. The method of Particle Swarm Optimization (MPSO) is a method that the authors apply. It is shown through various research here. PSO is very suitable combined with Support Vector Machine (SVM) (Danenas&Garsva, 2012), PSO and the Neural Network (Li et al, 2013), PSO with Self Organizing Map or SOM (o'neill&Brabazon, 2008). In this study, MPSO (Modified Particle Swarm Optimization) which is another variant of PSO was used to optimize the parameters of the Fk-NN. This study built a model to evaluate the provision of credit based on the classification of the Fk-NN and MPSO, in other words, the optimization of the parameters of the Fk-NN by MPSO is expected to improve the accuracy of classification.

2. TheoryStudy

2.1 Data Mining

Data mining contains the search trend or a desired pattern in a large database to help decision-making in time to come. Hopefully, the Data Mining is able to recognize these patterns in the data with minimal input (Hermawati:2013). Data mining explores the data base to find patterns that are hidden, looking for information to predict who may have been forgotten by analysts because it is located outside of their expectations (Alexander, 2009). Data mining can also take advantage of experiences or even mistakes in the past to improve the quality of the model and the results of the analysis, one with the ability of learning owned several data mining techniques such as classification and clustering (Kusnawi, 2007).

2.2 Fuzzy K Nearest-Neighbor

Fuzzy K-Nearest Neighbor (FK-NN) is one of the classification methods by combining the techniques of Fuzzy and K-NN. This method will explicitly predict the class followed by test data based on the comparison of the K closest.

Algorithm FK-NN gives the value of membership grade on the test data instead of putting the test data in a particular class. FK-NN is a classification method used to predict the test data using the value of the degree of membership of the test data in each class (Grace et al., 2018).

Before calculating the membership value on the Fuzzy K-NN, the first process is carried out using equation 1 below:

$$u_{ij} = \begin{cases} 0{,}51 + \left(\frac{n_j}{n}\right) * 0{,}49, if\ j = 1 \\ \left(\frac{n_j}{n}\right) * 0{,}49, if\ j \neq 1 \end{cases} \quad (1)$$

Description:

nj= Number of members of class j in a training data n

n= Number of training data used

j = a class of data

Next calculate the value of membership of each class with the equation 2 below:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

$$\frac{\sum_{j=1}^{k} u_{ij}\left(\|x-x_j\|^{-2/(m-1)}\right)}{\sum_{j=1}^{k}\left(\|x-x_j\|^{-2/(m-1)}\right)} \quad (2)$$

Description:
uij= the value of membership fuzzy on the example of testing (x, xj)
k= the value of the nearest neighbor
j = variable data membership data test
m = the weights that rank the magnitude of the m > 1.

2.3 PSO

Particle swarm optimization (PSO) is one of the methods to resolve the problem-the optimization problem are included in the meta-heuristic methods, meaning that PSO is associated with something that is random (stochastic) in solving the optimization problem faced. Some general terms commonly used in Optimization Particle Swarm can be defined as follows (Tuegeh, et al., 2009):

a) Swarm: the population of an algorithm.
b) Particle:member (individual) in a swarm.

Each particle represents a solution that potential problems are resolved. The position of a particle is determined by the representation of the solution at that time.

a) Pbest (Personal best): the position of Pbest of a particle which indicates the position of the particle which is prepared to get the best solution.
b) Gbest (Global best): position of the best particle in the swarm.
c) Velocity (vektor): a vector that drives the optimization process that determines the direction in which a particle is required to move (move) to fix the original position.
d) Inertia weight: inertia weight denoted w, this parameter is used to control the impact of the presence of velocity given by a particle.

This research was conducted in several stages, including the stage of pre-process data, the selection of the parameters k, m optimal use of MPSO and FKNN.

2.4 K-Fold Cross Validation (Evaluation Method Klasifikator)

K-fold cross validation starts by dividing the data a number of n-fold desired. In the process of cross validation the data will be divided into n partitions of equal size (D1, D2, D3, ...Dn), then the process of testing and training is performed n times. In iteration i, the partition will be the testing data and the rest will be the training data. The use of the number of fold best to test the validity, it is recommended menggunakan10-fold cross validation in the model (Prasad, 2014). The calculation accuracy by using equation 3 as follows:

$$\frac{\sum \text{test data correct classific ation}}{\sum \text{total test data}} \times 100\% \quad (3)$$

2.5 The Size Of The Similarity (Euclidean Distance)

According to Afrisawati (2013) to determine the correlation between two objects is by using the formula of Euclidean Distance as equation 4 below:

$d_{Euclidean}\ (x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

description:
d (xy)=distance data to the x to the center of the cluster y
x_i =the data to-i on the attributes of the data to n
y_i =data to-i on the attributes of the data to n

2.6 Data Normalization

Data attribute with a value that is large have a greater influence in the prediction of classification than the data with little value. To overcome such problems, the used normalization technique so that all the data is in the range of the same and no data has dominant influence to the other data. To calculate the normalization of the data used the formula with equation 5:

$$\text{normalization} = \frac{X - \min\ Data}{Maks\ Data - Min\ Data} \quad (4)$$

3 Research Methods

The purpose of classification in this research, namely to draft the best prediction model in classifying credit risk (smoothly or not smoothly/crash). For the framework of this study, the stages-stages are as follows:

a) Stages of first, select the dataset to be processed, then the data dipreprocessing that is normalized using the method of min-max.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

b) Further separation of the training data and testing data.
c) On the training data, performed the determination of the values m and k by using the method of MPSO.
d) The Value of m and k are selected is used as a parameter to classify the training data using the method FK-NN.
e) Performed the validation of the classification results using the method of k-fold cross validation to test the value of m and k selected in step 4.
f) Obtained values of m and k that have been validated.
g) Apply the value of m and k to classify the testing data using the method FK-NN which is expected to get the classification results with the best accuracy. To more clearly can be seen in figure 1 below:

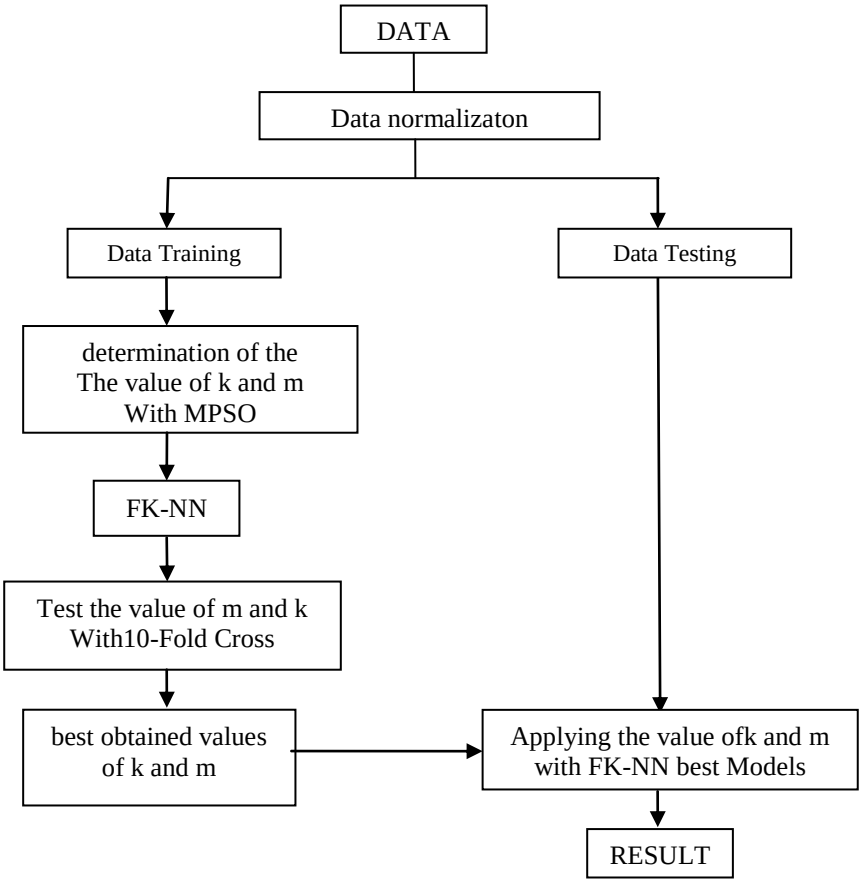

Fig 1 Research Framework

3.1 Data used

This research will evaluate the provision of credit based on the classification of the Fk-NN and MPSO using the dataset of the credit world published by the University of California Irivine (UCI) https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data) is a dataset of the German which has indeed been widely used in research assessment of credit for this. The characteristics of the data are described as follows:

Dataset
The Characteristics Of The Data Set : Multivariate
The Characteristics Of The Attributes : Categorial, Integer
A Lot Of Data
A Lot Of Attributes : 20
Missing values
Sector

3.2 Data Transformation

For data attributes that are categorical, it is necessary to be converted into numeric. The conversion of data categorical into numeric needs to be done because in the algorithm of the Fuzzy-KNN will do calculation of the distance using the formula Euclidean (Leidiyana, 2017). Pengubahannya can be done by replacing the data with a certain figure as long as it is consistent (Written, 2011). Suppose for attribute 1,

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

namely the Status of the account demand deposits held for which data is shaped categorical that:
A11 : ... < 0 DM
A12 : 0 ≤ ... < 200 DM
A13 : ... ≥ 200 DM / penugasan gaji untuk setidaknya 1 tahun
A14 : tidak ada akun giro
Nilai-nilainya diubah sebagai berikut:
A11 = 1
A12 = 2
A13 = 3
A14 = 4

4. Results and Discussion

Testing is performed on 1000 data credit risk, with comparison of the training data and test data that 10% of the 100 test data and 90% of the 900 training data. For the validation of the F-KNN was used 10-fold validation. Testing is done to obtain the value of m is best for the calculation of the Fuzzy and the value of the best k for KNN so as to maximize the accuracy of Fuzzy KNN.
The test scenario is to do 4 times the tests on the variation of the weight value of the inertia on the algorithm of MPSO, which is as follows:
Test 1 : weight of inertia: 0,5; weight cognitive: 1; weight social: 1
Test 2 : weight of inertia: 1; weight cognitive: 1; weight social: 1
Test 3 : weight of inertia: 1; weight cognitive: 2; weight social: 2
Test 4 : weight of inertia: 2; weight cognitive: 3; weights social: 3
The test results with the variation of the value of the weight inertia can be seen in the graph shown in figure 2 and the value of accuracy can be seen in table 2. This the results of the analysis of the determination of the testing 1, testing 2, testing 3, and testing 4 does not affect the accuracy only affects the speed of the particles in MPSO.

Table 2
Accuracy Testing with the variation of the value of the weight of the inertial

	Bobot inertia	Bobotkognitif	Bobotsosial	Akurasi
Pengujian 1	0.5	1	1	92,4%
Pengujian 2	1	1	1	92,4%
Pengujian 3	1	2	2	92,4%
Pengujian 4	2	3	3	92,4%

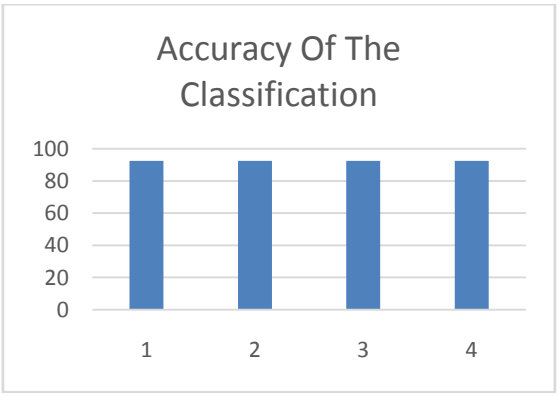

Fig 2 Classification accuracy on testing with variations the value of the weight of the inertial

Figure 3 is a display of the results of data processing using matlab applications. The Output of the application consists of 3 outputs, namely the best Accuracy for all iterations, the value of k is best (k_best) and the value of m the best (m-best). The movement of particles in the course of the algorithm MPSO can be monitored by observing the plot of which is shown as in figure 4. In figure 4 there are 50 particles (x) with the position expressed by the coordinates (k, m). for a particle located at coordinates (3,4) will test the solution accuracy for the value of k=3 and m=4. The results of testing the obtained value of k the best is 7

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

and the value of m is best 9. Obtained from testing the accuracy of 92.4%. The value of k=7 i.e. the number of neighbors for FKNN which produces the best accuracy. The value of m=9 in the process of classification using Fuzzy K-Nearest Neighbor effect on the value of the degree of membership of each test data against each class. The variable m is the weight of the rank is used to determine how much distance between neighbors when calculating the influence of neighbors on the value of membership

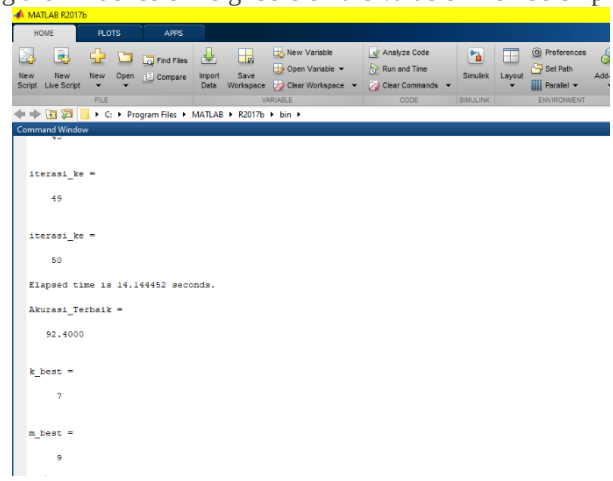

Fig 3 Display the Matla bresults of the MPSO F-KNN for 50 iterations

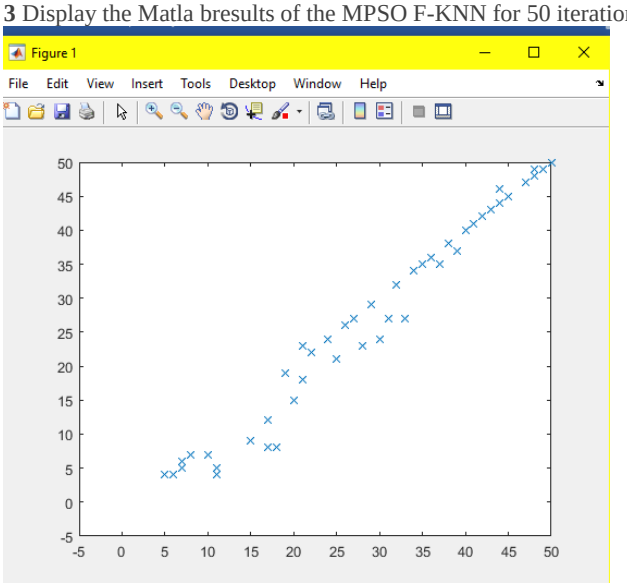

Fig 4 Plot of 50 in particles

5. Conclusions

The algorithm Fuzzy K-Nearest Neighbor (FK-NN) can be applied for the prediction of data classification of credit risk with an Algorithm Modified Particle Swarm Optimization (MPSO) to determine the value of k as the nearest neighbors and m as the weighting exponent of membership value for each class. There are two main processes in determining the classification of credit risk, namely data preprocessing, and classification process. The process of preprocessing aims to normalize the data of each attribute so each attribute can be influential in a balanced manner to the process of classification. Furthermore, the process of classification involves the stages, namely the search value k and the search value of m using the algorithm Modified Particle Swarm Optimization (MPSO) that each value of k and m found the particles sought the distance between the test data and training data, then look for the value of membership of each class in order to get the closest class which is the target class of the test data are new.

The level of accuracy in the Fuzzy K-Nearest Neighbor (FK-NN) with MPSO is influenced by several factors, among others, the value of k, and the value of m. On testing, FKNN accuracy is reached 92.4%, with the value of k is best at a value of 7 and the value of m is best in value 9.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

6. References

[1] Leidiyana,H.2013.Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit KepemilikanKendaraan Bermotor. *Jurnal Penelitian Ilmu Komputer, System Embedded & Logic*, Vol. 1(1), hal. 65-76

[2] Lopez & Jeronimo, J. 2015, *Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment*. A correlated-adjusted decision forest proposal. Expert Systems With Applications, 42 (13):5737-5753

[3] J. *Abellán*, G &*Castellano*, A comparative study on base classifiers in ensemble methods for ... Appl. 73, 1–10 (*2017*) 2. M. *Ala*'raj, M.F. *Abbod*, A new hybrid ensemble credit scoring model. Universidade de Brasília, Brasília, *2016*.

[4] Moula, A. K. A. 2015. Bank Credit Risk Analysis With K-Nearestneighbor Classifier: Case Of Tunisian Banks. *Accounting and Management Information Systems*, Vol. 14, No. 1, pp. 79-106.

[5] Kurama et, al 2015, *Behavior of Reinforced Concrete Beams with Recycled Concrete Coarse Aggregates*, Journal of Structural Engineering 141(3):B4014009, October2014

[6] Rosyid A, 2016. *Technological Pedagogical Content Knowledge: Sebuah Kerangka Pengetahuan Bagi Guru Indonesia di Era Mea*. Prosiding Seminar Nasional Inovasi Pendidikan Inovasi Pembelajaran Berbasis Karakter dalam Menghadapi Masyarakat Ekonomi ASEAN.

[7] Takyar, 2015, EphrinB2 signaling in osteoblasts promotes their differentiation by preventing apoptosis, The FASEB Journal 28:4482-4496, January 2015.

[8] 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.