Journal of Soft Computing Exploration

Improving Algorithm Accuracy K-Nearest Neighbor Using Z-Score Normalization and Particle Swarm Optimization to Predict Customer Churn

Muhammad Ali Imron [1], Budi Prasetiyo [2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 6, 2020
Revised Aug 20, 2020
Accepted Sept 7, 2020

Keywords:

Data Mining
German Credit
K-Nearest Neighbor
Particle Swarm Optimization
Z-Score

ABSTRACT

Due to increased competition in the business world, many companies use data mining techniques to determine the loyalty level of customers. In this business, data mining can be used to determine the loyalty level of customers. Data mining consists of several research models, one of which is classification. One of the most commonly used methods in classification is the K-Nearest Neighbor algorithm. In this study, the data which used are from German Credit Datasets obtained from UCI machine learning repository. The purpose of this study is to find out how Z-Score works to normalize the data and Particle Swarm Optimization to find the most optimal K value parameters, so the performance of the K-Nearest Neighbor algorithm is more optimal during the classification. The methods which were used to normalize the data are Z-score and Particle Swarm Optimization to determine the most optimal K value. The classification was tested using confusion matrix to determine the generated accuracy. From the finding of this study, the application of Z-score normalization and Particle Swarm Optimization with the K Nearest Neighbor algorithm succeed in increasing the accuracy up to 14%. The initial accuracy was 68.5%, and after applying the normalization of Z-Score and Particle Swarm Optimization, the accuracy became 82.5%.

Corresponding Author:

Muhammad Ali Imron
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: aliimron@students.unnes.ac.id

1. INTRODUCTION

Data era is an era where data which rapidly developed, widely distributed and have large capacities require an appropriate and organized processing method so that it can be maximally utilized [1]. From the available data, its' information will be extracted and was expected to be applied into larger data that has never been known before.

In the business world, customers are the main asset. Therefore, various ways have been taken by companies, so customers do not stop to subscribe [2]. To get a new customer costs up to 10 times more than the cost of retaining the existing customers. The high cost to get new customers, of course, companies prefer to retain the existing customers. Based on that fact, many companies turn to retain the existing customers and avoid customer churn [3]. Based on increasing competition and supply in the industrial market, many companies are utilizing data mining to predict [4]. Data mining is an activity used to find interesting patterns for large amounts of data [5].

There are several stages in data mining, those are pre-processing, processing, and post-processing stage. The pre-processing stage consists some stages, such as data cleaning, data integration, data reduction, and data transformation [6]. In the data transformation consists of some methods to process the transformation, such as smoothing, generalization, normalization, aggregation and attribute construction [7]. According to Junaidi, et al. normalization is the process by which a numerical attribute is mapped or scaled within a certain range. Data normalization is useful to minimalize data refraction in data mining because the values of attribute in data usually have different ranges [8].

There are several normalization techniques which often be used, those are min-max normalization, Z-Score normalization and decimal scaling normalization, all of which have the goal of mapping data to a certain scale [9]. Z-Score normalization is data normalization used to provide data ranges using mean and standard deviation [10].

To accomplish optimization problems, Particle Swarm Optimization (PSO) algorithm is one of the meta-heuristic algorithms that commonly used [11]. In some cases, it has been proven that PSO is more competitive [12]. This optimization method is proven effective and succeed to be used to solve multidimensional and multi-parameter optimization problems in machine learning such as neural networks and classification techniques algorithms [13].

K-Nearest Neighbor (KNN) algorithm is a method to classify objects based on learning data which has the closest distance to the object. This technique is very simple and easy to implement. It is similar to clustering technique, which is grouping to the new data based on the distance of the new data to several data / nearest neighbors. Before searching for the distance of the data to the neighbors, we have to determine the value of K neighbor.

2. METHOD

Data processing stage in this study was carried out in several stages, starting from converting the data from .data to .csv extension, the data obtained from UCI machine learning repository had its available numerical data already, thus the transformation of the data from nominal to numeric was not needed. The next stage was normalization of Z-Score and data mining stage. For more details about the methods used in this study, it can be seen in Figure 1.

2.1 Z-score

Certain in the data mining process. Whereas, according to [7]. Normalization is one of the data transformation processes in the data mining process where numerical attributes are scaled in a smaller range. The Z-Score value ranges between infinite negative and positive numbers. Different from the normalized values, the Z-Score does not have a minimum and maximum value set [14]. There are several methods that are usually applied in data normalization, those are: min-max normalization, Z-Score normalization and normalization by decimal scaling. Z-Score normalization is a method of normalizing data when the range of data is not known with any certainties, thus it is necessary to calculate the range using the mean and standard deviation of the data [10].

The application of Z-Score normalization stage in data mining pre-processing process is as follows:

Step 1: Find the average value of each numeric attribute.

Step 2: After found the value of mean or average of each attribute, the next step is look for data variance from the numeric attributes. Variances are used to find out how far the data spread from the mean. Low variance shows that data was clustered very close around the mean and vice versa. In calculating the data variance, each initial value of the data in numeric attribute reduced by the mean value of each attribute. After found the value of the mentioned reduction, the next step is to square the reduction results then add the squared results for each attribute. After found the sum result of each attribute as above, the next step is to divide the aforementioned sum result by (n), for n is the number of the data records and will produce a data variance value for each numeric attribute.

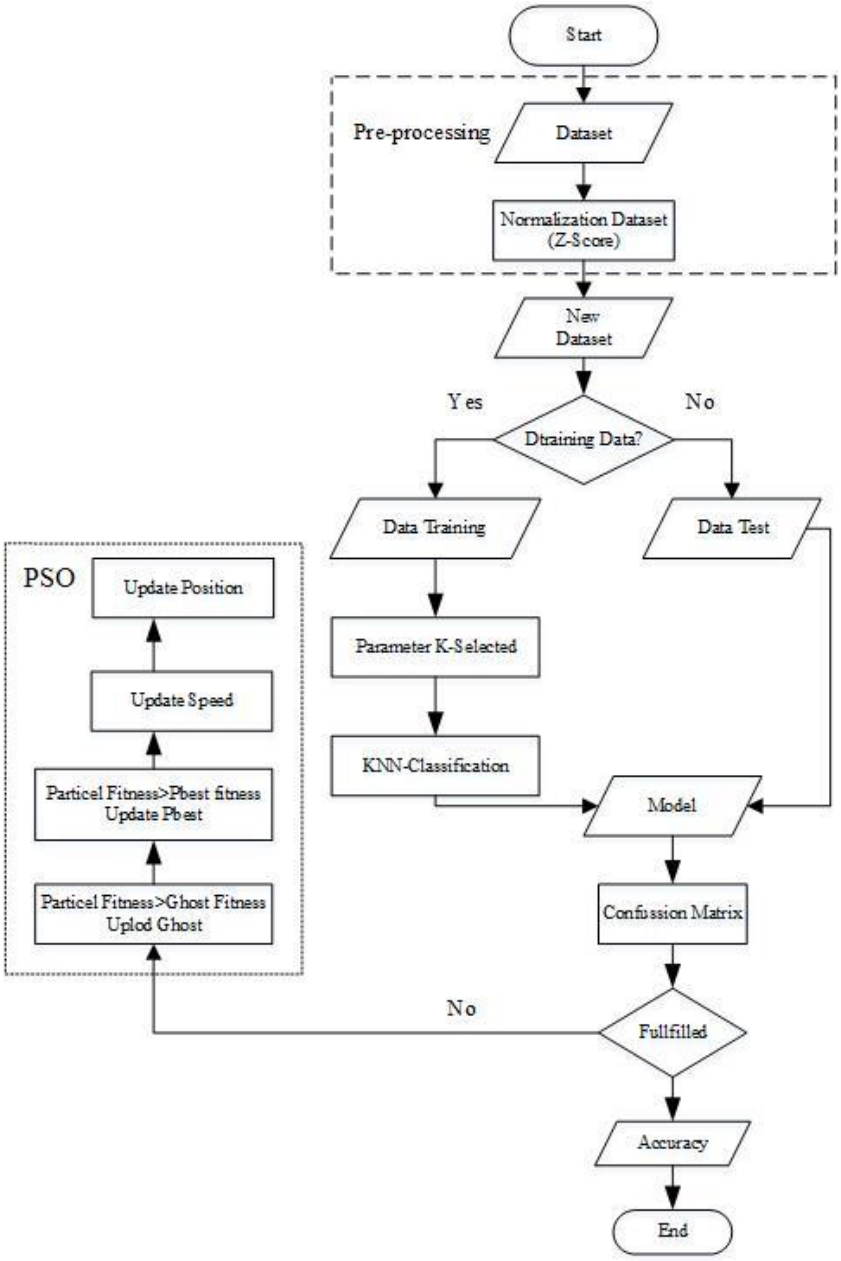

Figure 1. Flowchart of method

Step 3: After found the value of the data variance, the next step is look for the standard deviation value. The value of the standard deviation can be found by calculating the square root of the data variance value of each numeric attribute. To find the standard deviation value, we can use the Equation 1.

$$SD_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}, \quad (1)$$

Information:

SD_x = Standard Deviation
n = Number of Samples
x = Average
x_i = Value of x to i

Step 4: The next step is calculate the Z-Score normalization value with Equation 2.

$$Z = \frac{X - \bar{X}}{SD_x} \quad (2)$$

Information:

Z = Normalization results
x = Value to be normalized
$\bar{X}$ = Average Value
SD_X = Standard Deviation

2.2 Partical Swam Optimization

Particle Swarm Optimization (PSO) is one of the basic techniques of the swarm intelligence system to solve optimization problems in the search for space as a solution. This optimization method has been proven effective and has been successfully used to solve multidimensional and multi-parameter optimization problems [15]. The stages of PSO in optimizing can be shown as in Figure 2.

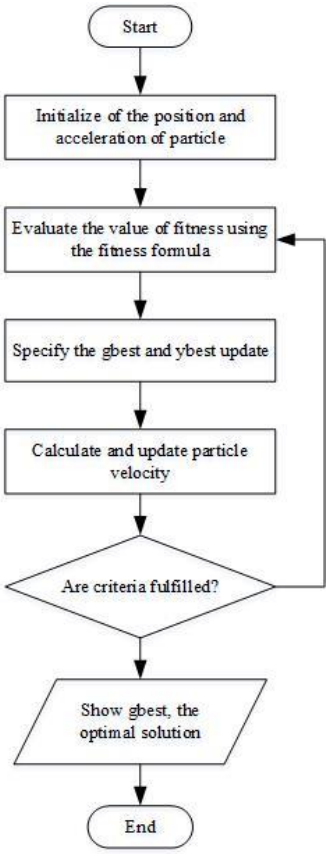

Figure 2 Particle Swarm Optimization

Based on the Figure 2, it can be seen more clearly PSO steps as follows.

Step 1: Initialize particle position (*xt*), weight of inertia (*w*) = 0.72, and acceleration coefficients (*c1* and *c2*) = 0.7. Initialize particle velocity (*vt*) = 0. Number of particles = 50 and iterations performed = 100.

Step 2: Calculate and evaluate the fitness value of each particle by using the KNN algorithm.

Step 3: Determine the *p*best value of each particle based on the accuracy value produced by KNN. Determine *g*best based on the highest *pbest* value.

Step 4: Calculate the velocity and position of the particle by using Eqs. 3 and 4.

$$v_{id}^{t+1} = w \times v_{id}^{t} + c_1 \times r_{1i} \times (p_{id} - x_{id}^{t}) + c_2 \times r_{2i} \times \left(p_{gd} - x_{id}^{t}\right) \quad (3)$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \quad (4)$$

Step 5: Show *gbest* and optimal solutions with the best K value which will be used.

2.3 K-Nearest Neighbor Algorithm

K-Nearest Neighbor is a classification algorithm which remains consistent in a large amount of data and classifies based on the closest distance between the data evaluated by the closest point in the training data. The KNN algorithm is more flexible because it is based on the proximity of existing training data [16]. The KNN stage is carried out in the training data during the classification process which can be seen in Figure 3.

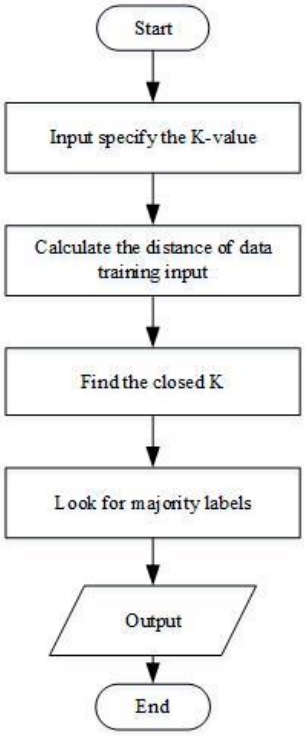

Figure 3 . K-Nearest Neighbor

According to the Figure 3, the steps of the KNN algorithm can be seen more clearly.

Step 1: Defining the value of K.
Step 2: Calculating the distance or Euclidean values between the testing data and the training data.
Step 3: Grouping the data based on the distance calculation or Euclidean.
Step 4: Grouping the data based on the distance calculation or Euclidean.
Step 5: Selecting the class that most emerges from the number of the selected K to be used as a prediction result.

The training data on attribute 1 can be seen in Equation 5.

$$X_1 = (X_{11}, X_{12}, \ldots, X_{1n}) \tag{5}$$

The training data on attribute 2 can be seen in Equation 6.

$$X_2 = (X_{21}, _{22}, \ldots, X_{2n}) \tag{6}$$

Whereas, to find the Euclidean distance expressed by Equation 7.

$$d(X_1, X_2) = \sqrt{\sum_r^n}(a_r(X_1) - a_r(x_{12}))^2 \tag{7}$$

2.4 Evaluate using Confusion Matrix

The evaluation stage was carried out at the end of the study process. This stage was useful to test the model and to calculate the accuracy result. In this study, the evaluation was carried out using confusion matrix. The steps are as follows.

Step 1: Enter the test results to the confusion matrix table as in Table 1.

Table 1. Confusion matrix Testing

Actual	**Predicted**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Step 2: Calculate the accuracy value, determine the highest accuracy with Equation 9.

$$Accuracy = \frac{TP+TN}{P+N} \text{ x } 100\% \quad (9)$$

Step 3: State the conclusions from the accuracy results obtaine

3. RESULT AND DISCUSSION

This research was conducted using tools, namely: Sublime Text 3 text editor, Python 3 programming language, skicit-learn library, as well as the documentation pyswarms. While, the material which was used is German Credit Data obtained from the UCI Machine Learning Repository.

Z-Score normalization method can increase accuracy by representing the original data into new data with almost similiar range of values and a narrow range of values. This simplify the data, so that the data mining process can be more optimal and increase accuracy for the KNN algorithm by 80,5%.

The PSO stage was used to optimize the K parameter in KNN algorithm. Each iteration will obtain its' best position with the lowest cost value that called as the best cost. After doing all the iterations, the cost value of each iteration was compared to obtain the final best cost. This final best cost will be used as a recommendation of the optimization process. This recommendation is in the form of the lowest cost value and the chosen K value which can produce the an accuracy of 73,5%.

This study recorded every accuracy results from the classification process that had been done. The results can be seen in Table 2.

Table 2 Results of Each Method Used

Algorithm	Accuracy
KNN	68,5%
KNN+*Z-Score*	80,5%
KNN+PSO	73,5
KNN + *Z-Score* + PSO	82,5%

According to the Table 2, it is seen the improvement of each method used. KNN algorithm without using the normalization of Z-Score and PSO was 68.5%. KNN algorithm with Z-Score normalization without using PSO produces an accuracy of 80.5%. KNN algorithm with PSO without using Z-Score normalization produces an accuracy of 73.5%. Purposed method which in this case is the KNN algorithm using Z-Score and PSO normalization produces accuraty of 82.5%. Thus, it can be concluded that there is an increase in the accuracy of 14% by comparing the KNN algorithm without using the normalization of Z-Score and PSO with the Purposed method in this study.

Table 3 Research Accuracy Results

Method	Accuracy
Sobran *et al*	65,3%
Safitri & Muslim	74,9%
Jeatraku *et al*	77,9%
Purposed Method	82,5%

The method used in this study was compared to the previous studies, it can be seen that the accuracy generated in this study is better than some previous studies using German Credit Data as in the Table 3.

4. CONCLUSION

From the finding and discussion of this study related to the implementation of the normalization of Z-Score and PSO in order to improve the accuracy of KNN algorithm using German Credit Datasets obtained from the UCI Machine Learning Repository, it can be concluded that the application of normalization Z-Score can provide a range of each value in the attribute on German Credit Datasets, thus it is increasing the accuracy of the KNN algorithm. Then, the application of PSO on German Credit Datasets was used to find the best K parameter value, after the optimization results obtained, the best K parameter value was classified using the KNN algorithm. The accuracy results obtained on the application of the KNN algorithm using normalization Z-Score and PSO is 82.5%. The increase of the accuracy is 14% from the application of the KNN algorithm which only has an accuracy of 68.5% for German Credit Datasets objects originating from the UCI repository of machine learning.

REFERENCES

[1] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques Third Edition. USA: Elsivier. 2012

[2] Y. Liu, and Y. Zhuang, "Research model of churn prediction based on customer segmentation and misclassification cost in the context of big data, " *J. of Comp. & Comm.* vol. 03, pp. 87-93, 2015.

[3] Y. Huang and T. Kechadi, " An effective hybrid learning system for telecommunication churn prediction, " Exp. Sys. with Appl. Vol. 40, pp. 5635-5647, 2013

[4] I. Brandusoiu, and G. Toderean, "Churn Prediction in the Telecommunications Sector using Support Vector Machines, "*Ann. of the Oradea University*, vol. 22, no. 1, pp. 19–22, 2013.

[5] E. Sugiharti, S. Firmansyah, and F. R. Devi, "Predictive Evaluation of Performance of Computer Science Students of Unnes Using Data Mining Based on Naïve Bayes Classifier (NBC) Algorithm, " *J. of Theoretical & Appl. Info. Tech.*. , vol. 95, no. 4, pp. 902–911, 2017.

[6] P. Plawiak, M. Abdar and U. R. Acharya, "Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring, " *App. Soft. Compt*., vol. 84, pp. 105740, 2019.

[7] C. Ordonez, S. Maabout, D. S. Matusevich, and W. Cabrera, "Extending ER models to capture database transformations to build data sets for data mining, " *Data & Know. Eng*. vol. 89, pp. 38-54, 2014.

[8] X. Zhong, and D. Enke, "A comprehensive cluster and classification mining procedure for daily stock market return forecasting, " *Neurocomputing*, vol. 267, pp. 152-168, 2017.

[9] C. Saranya, and Munikandan, "A Study on Normalization Techniques for Privacy Preserving Data Mining, " *Int. J. of Eng. & Tech.*, vol. 5, no. 3, pp. 2701, 2013

[10] H. Goyal, Sandeep, Venu, R. Pokuri, S. Kathula S and N. Battula, " Normalization of Data in Data Mining, " *Int. J. of Soft. & Web Science,* vol. 10, no. 1, pp. 32-33. 2014.

[11] L. A. Ashari, M. A. Muslim, and Alamsyah, "Comparison Performance of Genetic Algorithm and Ant Colony Optimization in Course Scheduling Optimizing, " *Sci. J. of Info.* vol. 3, no. 2, pp. 149-158, 2016.

[12] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, " Optimization of C4.5 Algorithm-based Particle Swarm Optimization for Breast Cancer Diagnosis " *J. of Physic*, vol. 983, no. 1, pp. 1-5, 2017

[13] S. W. Fei, M. J. Wang, Y. B. Miao, J. Tu, and C. L. Liu, "Particle Swarm Optimization-based Support Vector Machine for Forecasting Dissolved Gases Content in Power Transformer Oil, " *Energy Conversion and Manag.*, vol. 50, no. 6. pp. 1604-1609, 2009.

[14] A. Pandey, and A. Jain, "Comparative Analysis of KNN Algorithm using Various Normalization Techniques " *Int. J. of Comp. Net. & Infor Sec.* vol. 11, no. 04, pp. 36-42, 2017.

[15] Sumathi, S., & Surekha, P. Computational Intelligence Paradigms: Theory and Applications Using Matlab (1st ed.). Boca Raton: CRC Press. 2009

[16] M. R. Hidayah, I. Akhlis, and E. Sugiharti, "Recognition Number of The Vehicle Plate Using Otsu Method and K-Nearest Neighbour Classification , " *Journal of Soft Computing Exploration*, vol. 4, no. 1, pp. 66-74. 2017.