Analysis Accuracy Of Forecasting Measurement Technique On Random K-Nearest Neighbor (RKNN) Using MAPE And MSE

S Prayudani¹, A Hizriadi², Y Y Lase³, Y Fatmi⁴, Al-Khowarizmi⁵

¹,³,⁴Jurusan Teknik Komputer dan Informatika, Politeknik Negeri Medan, Medan, Indonesia
²Fakultas Ilmu Komputer dan Teknologi Informatika, Universitas Sumatera Utara, Medan, Indonesia
³Fakultas Teknik, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

E-mail: santiprayudani@polmed.ac.id

Abstract. Forecasting is apply because of complexity and uncertainty faced by high-dimensional data available in the fields of bioinformatics, chemometrics, banking and other applications. A process for systematically estimating what is most likely to happen in the future based on past and present data requires an appropriate forecasting model, so that the difference between what happens and the estimated results can be minimized. To get the right method, a measuring technique is needed to detect the accuracy of forecasting value. In this paper we discuss the technique of measuring forecasting accuracy with Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE) using the Random K-Nearest Neighbor (RKNN) method. With the two measuring technique for the horizontal modeling above, the smallest MSE and MAPE values are chosen (the smallest error value). From the results of the analysis of the calculation of forecasting accuracy measurement values during training with RKNN, the MAPE accuracy value is 0.728427% and MSE is 0.545751, while the smallest accuracy value is achieved using MSE which is 0.545751.

1. Introduction

Currently high-dimensional data in various fields such as bioinformatics, chemometrics, banking and others. High-dimensional data processing is very challenging for users, including a small number of observations $n$ but a large number of independent variables $p$, the existence of irrelevant variables, and the difficulty of making model prediction directly from existing data [6]. Various methods may be applied in overcoming the problems of high-dimensional data. One of the methods in this case is K-Nearest Neighbor (KNN) [6].

However, dealing with high-dimensional data, a process is needed to estimate what is most likely to occur in the future based on past data series and require an appropriate forecasting model, so the KNN method was developed by various researchers using Random KNN (RKNN). One of the advantages of RKNN is that missing values can be easily calculated. RKNN leads to significant performance both in terms of computational complexity and classification accuracy [5]. Every forecasting requires a forecasting accuracy measurement tool to get the smallest error value in forecasting [9].
Alamgir & Luxburg conducted a study with RKNN by searching for the shortest distance from the graph so that the results with the shortest distance were influenced by the weight values found in the KNN [1]. In his research, he did not use any forecasting accuracy measurement technique because it focused on distance.

In 2011 Sungkawa & Megasari research with different methods applying a measure of the accuracy of the time series data forecast values in the sales volume forecasting model selection at PT. Satriamandiri Citramulia. There is a measuring technique using mean square error (MSE) and mean absolute percentage error (MAPE) [10].

In 2018, related to KNN as done by Satriya, Santoso & Sutrisno conducted research with Ensemble KNN which was the same as the results of the development of the KNN predicting the value of the rupiah against the US dollar and the final results there were forecasting accuracy measures from the test results using MAE buy with the results of 456.56, MAE with the results of 460.96, MAPE buying with the results of 3.47%, MAPE selling with the results of 3.47%, RMSEP buying with the results 534.88, and RMSEP selling with the results of 540.07 [2].

In this study, we discuss the accuracy of measuring forecasting technique on the RKNN method by comparing forecasting accuracy measurements, which is MAPE and MSE, and then comparing its result. The sample data gets from dataset of UCI Machine Learning, which is dataset of bank credit approval, consist of 5 attributes and 10,000 data samples.

2. Method
The general architecture of this study can be seen in Figure 1.

![Figure 1. General Architecture](image)

The dataset that has been taken is stored in the database. And then enter all old data. Input new data to get proximity value. Data is normalize before processing data training to have a range from 0 to 1. Calculate the value of the proximity of the training data. Calculate all old data closeness (similarity) with new data. Displays all old data that has the proximity to new data. Calculate the value of forecasting accuracy with MAPE and MSE. Comparison of the two forecasting accuracy values.

2.1. Random K-Nearest Neighbor (RKNN)

Random K-Nearest Neighbor (RKNN) is designed specifically for the classification of high-dimensional datasets, the development of the K-Nearest Neighbor (KNN) algorithm [6]. KNN algorithm is an approach to look for cases by calculating the closeness between new cases and old cases based on matching weights of a number of features that have similarity [4]. The similarity of functions (similarities) will produce a value that determines whether there are new cases with cases that exist on a case basis. To determine the similarity can be done with the similarity of the euclidean distance function. Therefore, distance is often normalized by dividing the size for each attribute with a range (i.e. the maximum value - minimum value) attributes that are valuable for each attribute have a normalized new range of 0 to 1.
\[ y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1) \]

Where \( x \) is data value, \( y \) is normalization value, \( x_{\min} \) is minimum value as in 0 (zero), \( x_{\max} \) is maximal values as in 1 (one). After normalization, the data then calculates the proximity value. This calculation process is used in finding predictions. The equation formula for calculating proximity between two cases is as follows:

\[
\text{similarity}(T, S) = \frac{\sum_{i=1}^{n} f(x_i, x_i) \times w_i}{w_i} \quad (2)
\]

Where \( T \) is new case, \( S \) is cases that in storage (old cases), \( n \) is number of attributes in case, \( i \) is individual attribute between 1 to \( n \), \( f \) is similarity attribute function between case \( T \) (new case) and case \( S \) (old case).

2.2. The Mean Squared Error (MSE)

The Mean Squared Error (MSE) is method for evaluating forecasting methods. Each error or remainder is squared. Then added up and added to the number of observations. This approach regulates large forecasting errors. The method produces moderate errors which might be better for small errors.

\[
MSE = \frac{\sum_{t=1}^{n} (a - b)^2}{n} \quad (3)
\]

Where \( a \) is actual data, \( b \) is result data, \( n \) is amount of data

2.3. The Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error (MAPE) is calculated using absolute errors in each period divided by the actual observation values for that period. Then, the average absolute percentage error. This approach is useful when the size or size of predictive variables is important in evaluating the accuracy of predictions. MAPE indicates how much error in forecasting is compared to the real value in the series. MAPE can also be used to compare the accuracy of the same or different methods in two different series and measure the accuracy of the estimated value of the model expressed in terms of the absolute percentage error average.

\[
MAPE = \frac{\sum_{t=1}^{n} |a - b|}{n} \times 100\% \quad (4)
\]

Where \( a \) is actual data, \( b \) is result data, \( n \) is amount of data

2.4. Analysis of the Random K-Nearest Neighbor (RKNN) Method

Looking for a solution for a new bank customer (new case) by using a solution from a previous bank customer (old case). To find out which previous bank customers to use, the closeness of the case of new customers is calculated with all cases of old customers. The case of old customers with the greatest closeness will be taken for use in the case of new customers. To predict whether a new bank customer has a problem or not based on the data owned.

**Table 1. Old Customer Case Example**

| Customer Name | Education | Sex | Income (Monthly) | Expense (Monthly) | Job          | Problem |
|---------------|-----------|-----|------------------|------------------|--------------|---------|
| Dedi          | >=S1      | M   | 2jt – 3jt        | <=2jt            | Entrepreneur | N       |
| Budi          | >=S1      | M   | 2jt-3jt          | 2jt-3jt          | Private Employee | Y      |
| Linda         | D1-D3     | F   | >=3jt            | 2jt-3jt          | Civil Servant | N       |

Determine the weight of each attribute.
Table 2. Weight of Each Attribute

| Attribute     | Weight |
|---------------|--------|
| Education     | 0.5    |
| Sex           | 0.5    |
| Income/Bulan  | 1      |
| Expense/Bulan | 1      |
| Job           | 0.75   |

Determine the proximity values in the attribute

- proximity value of education attribute

Table 3. Proximity Value of Education Attribute

| Education   | Education   | Proximity |
|-------------|-------------|-----------|
| <=SMA       | <=SMA       | 1         |
| D1-D3       | D1-D3       | 1         |
| >=S1        | >=S1        | 1         |
| <=SMA       | D1-D3       | 0.5       |
| D1-D3       | <=SMA       | 0.5       |
| <=SMA       | >=S1        | 0.4       |
| >=S1        | <=SMA       | 0.4       |
| D1-D3       | >=S1        | 0.75      |
| >=S1        | D1-D3       | 0.75      |

- proximity value of sex attribute

Table 4. Proximity Value of Sex Attribute

| Sex | Sex | Proximity |
|-----|-----|-----------|
| M   | M   | 1         |
| F   | F   | 1         |
| M   | F   | 0.7       |
| F   | M   | 0.7       |

- proximity value of income attribute

Table 5. Proximity Value of Income Attribute

| Income  | Income  | Proximity |
|---------|---------|-----------|
| <=2jt   | <=2jt   | 1         |
| 2jt-3jt | 2jt-3jt | 1         |
| >=3jt   | >=3jt   | 1         |
| <=2jt   | 2jt-3jt | 0.4       |
| 2jt-3jt | <=2jt   | 0.4       |
| <=2jt   | >=3jt   | 0.75      |
| >=3jt   | <=2jt   | 0.75      |
| 2jt-3jt | >=3jt   | 0.5       |
| >=3jt   | 2jt-3jt | 0.5       |
• proximity value of expense attribute

**Table 6. Proximity Value of Expense Attribute**

| Expense | Proximity |
|---------|-----------|
| <=2jt   | <=2jt     | 1   |
| 2jt-3jt | 2jt-3jt   | 1   |
| >=3jt   | >=3jt     | 1   |
| <=2jt   | 2jt-3jt   | 0.4 |
| 2jt-3jt | <=2jt     | 0.4 |
| <=2jt   | >=3jt     | 0.75|
| >=3jt   | <=2jt     | 0.75|
| 2jt-3jt | >=3jt     | 0.5 |
| >=3jt   | 2jt-3jt   | 0.5 |

• proximity value of job attribute

**Table 7. Proximity Value of Job Attribute**

| Job            | Proximity |
|----------------|-----------|
| Entrepreneur   | 1         |
| Private Employee | 1       |
| Civil Servant  | 1         |
| Entrepreneur   | 0.4       |
| Private Employee | 0.4    |
| Entrepreneur   | 0.6       |
| Civil Servant  | 0.6       |
| Private Employee | 0.3   |
| Civil Servant  | 0.3       |

Settlement of cases, which is there are new customers with the following attribute values:

- **Education**: D1-D3
- **Sex**: M
- **Income Monthly**: 2jt-3jt
- **Expense Monthly**: < 2jt
- **Job**: Entrepreneur

**Determine the weight of each attribute and calculating Proximity between new customer case with Dedi’s case.**

**Table 8. Proximity New Case with Dedi**

| Attribute | New Case | Old Case | Proximity Value | Attribute Weight |
|-----------|----------|----------|-----------------|------------------|
| Education | D1-D3    | >=S1     | 0.75            | 0.5              |
| Sex       | F        | M        | 0.7             | 0.5              |
| Income    | 2jt-3jt  | 2jt – 3jt| 1               | 1                |
| Expense   | < 2jt    | <=2jt    | 1               | 1                |
| Job       | Entrepreneur | Entrepreneur | 1 | 0.75 |

- Proximity new customer case with Dedi’s case :
  \[ \frac{0.75 \times 0.5 + 0.7 \times 0.5 + 1 \times 1 + 1 \times 1 + 1 \times 0.75}{0.5 + 0.5 + 1 + 1 + 0.75} = \frac{3.475}{3.75} = 0.926667 \]
• Determine the weight of each attribute and calculating Proximity between new customer cases with Budi’s case

| Attribute       | New Case | Old Case | Proximity Value | Attribute Weight |
|-----------------|----------|----------|----------------|------------------|
| Education       | D1-D3    | >=S1     | 0.75           | 0.5              |
| Sex             | F        | M        | 0.7            | 0.5              |
| Income          | 2jt-3jt  | 2jt-3jt  | 1              | 1                |
| Expense         | < 2jt    | 2jt-3jt  | 0.4            | 1                |
| Job             | Entrepreneur | Private Employee | 0.4 | 0.75 |

• Proximity new customer case with Budi’s case:

\[
\text{Proximity} = 0.75 \times 0.5 + 0.7 \times 0.5 + 1 \times 1 + 0.4 \times 1 + 0.4 \times 0.75 = 2.425 \times 3.75 = 0.646667
\]

From the proximity calculation results between the new case with Dedi’s, Budi’s and Linda’s cases, it found that the biggest proximity value gets in Dedi’s case, so the classification in Dedi’s case is used, which is the new customer would not have a problems.

2.5. Analysis of Accuracy Forecasting Measurement Technique

Based on the case of subchapter 2.3, the calculation of value accuracy forecasting measurement technique is carried out when training using RKNN. The results of training using RKNN can be seen from the following table 10.

| Customer Name | Education | Sex | Income (Monthly) | Expense (Monthly) | Job              | Problem | Proximity Value |
|---------------|-----------|-----|------------------|-------------------|------------------|---------|----------------|
| Dedi          | >=S1      | M   | 2jt - 3jt        | <=2jt             | Entrepreneur     | N       | 0.926667       |
| Budi          | >=S1      | M   | 2jt-3jt          | 2jt-3jt           | Private Employee | Y       | 0.646667       |

MAPE calculation:

\[
MAPE = \frac{\left( \frac{1 - 0.926667}{0} \right) + \left( \frac{1 - 0.646667}{1} \right)}{2} \times 100\% = 0.117778\%
\]

The MSE calculation:

\[
MSE = \frac{(1 - 0.926667)^2 + (1 - 0.646667)^2}{2} = 0.1348
\]

3. Result and Discussion

At training stage we analyze accuracy of forecasting measurement technique on RKNN Method. From 10000 number of old data case, the proximity value in training will be sought by using MAPE and MSE values on new data case in RKNN. The display of the system is shown in figure 2 and figure 3.
4. Conclusion
Based on the testing training stage the accuracy of the RKNN method with MAPE and MSE on 10000 old data case, the calculation using RKNN is get 0.728427% for MAPE and 0.545751 for MSE. We get the smallest error accuracy value is using MSE, which is 0.545751. It shown in table 12.

Table 11. MAPE and MSE Calculation Results

| Measurement Technique | Value    |
|-----------------------|----------|
| MAPE                  | 0.728427%|
| MSE                   | 0.545751 |

References
[1] Alamgir, M., & von Luxburg, U. (2012). Shortest Path Distance in Random K-Nearest Neighbor Graphs. Retrieved from http://arxiv.org/abs/1206.6381
[2] Hary, R., Satriya, D., & Santoso, E. (2018). Implementasi Metode Ensemble K-Nearest Neighbor untuk Prediksi Value Tukar Rupiah Terhadap Dollar Amerika, 2(4), 1718–1725.
[3] Heizer, J. & Render, B. (2005). Operations Management: Manajemen Operasi. Salemba Empat:
[4] Kusrini, Hartati. S, Wardoyo. R & Harjoko. A. 2009. “Perbandingan Metode Nearest Neighbor Dan Algoritma C4.5 Untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Di Stmik Amikom Yogyakarta”. Jurnal Dasi. Vol 10. No 1. Pp. 114-132.

[5] Ishak, A. (2010). Manajemen Operasi, Edisi Pertama. Yogyakarta: Graha Ilmu

[6] Li, S. (2009). Random KNN Modeling and Variable Selection for High Dimensional Data, Disertasi, Lane Departement of Computer Science & Electrical Engineering West Virginia University.

[7] Li, S., Harner, E. J., & Adjeroh, D. A. (2011). Random KNN feature selection - a fast and stable alternative to Random Forests. BMC Bioinformatics, 12(1), 450. https://doi.org/10.1186/1471-2105-12-450

[8] Li, S., dkk. (2015). Package "rknn"

[9] Noi, P. T., & Kappas, M. (2018). Comparison of Random Forest, K-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. Sensors (Switzerland), 18(1). https://doi.org/10.3390/s18010018

[10] Sukma, A., Ramadhan, D., Santoso, B. P., Sari, T. R., & Wiraswari, N. M. A. K. (2014). K-Nearest Neighbor Information Retrieval. Retrieved from http://web.unair.ac.id/admin/file/

[11] Sungkawa, I., & Megasari, R. T. (2011). Value Ramalan Data Deret Waktu Dalam Seleksi Model Peramalan Volume Penjualan PT Satriamandiri Citramulia. ComTech, 2(2), 636–645.

[12] Widiarsana, O., Putra, N.W., Budiyasa, P.G.I., Bismantara, A.N.I., Mahajaya, S.N. (2011). Data Mining: Metode Clasification K-Nearest Neighbor (KNN). Bali: Program Studi Teknologi Informasi Universitas Udayana.