K-nearest neighbor (KNN) with global GINI diversity index for classification subsidy food in Semarang city, Indonesia

D Ispriyanti, A Prahutama, Mustafid

1Department of Statistics, Faculty of Science and Mathematics, Diponegoro University Jl. Prof Soedharto, SH, Tembalang, Semarang 50275, Indonesia

Corresponding author: dwiispriyanti@yahoo.com

Abstract. K-Nearest Neighbor (K-NN) is one of the classification methods using the distance from the variable. Determination of the distance used greatly affects the results of the classification. The distance commonly used as a weighting is the Euclidean distance. The Euclidean distance does not distinguish between categories and variables, this is certainly a problem in itself. Therefore the weighting used is the global GINI diversity index. This method will be applied to the classification of subsidy food in Semarang city, based on the factors that influence it. Independent variables that are used include the field of business of household heads (X1); Employment Status (X2); Homeownership status (X3); Roof Building Materials (X4); Main wall material of the house (X5); Main Material of the house floor (X6); Drinking Water Source (X7) and Main Cooking Fuel (X8). Based on the results of classification using the KNN method with a global weighting GINI diversity index, with training: testing is 80: 20 obtained an accuracy of 85% with a value of K = 5.

1. Introduction
Classification is one of the techniques in machine learning that is classified into supervised learning. Classification provides an analysis of the factors that influence class labeling. From the results of classification analysis, we can predict the class label of an object that has certain characteristics (independent variables). Some classification methods with supervised learning include K-Nearest Neighbor (KNN), Fuzzy K-NN (FKNN), FKNN in Every Class (FKNNC), CART, C-45 [1]. KNN is a classification method based on the distance between locations. KNN concept calculates distances using Euclidean distances. Euclidean distance can interpret distances well for numeric type data. As for the Euclidean distance categorical data type, it is less interpreting. Therefore the distance calculation for categorical data uses Simple Matching Coefficient (SMC). While the weighting used is the global GINI diversity index. According to Chen and Guo (2015) [2], the weight given to each variable indicates the size of a variable contributing to class labeling. The greater the contribution, the greater the weight.

Subsidy food (acceptance of rice usually called “beras miskin”) is a government program that aims to improve the welfare of the community. This program is an implementation of the Medium-Term Development Plan (RPJM) [3]. This is also the implementation of the government in implementing the Sustainable Development Goal (SDG’s), and one of the programs in reducing poverty. Currently, SDG’s was developed into the Millennium Development Goals (MDG’s). In the MDG program, one of the goals is to reduce poverty and eliminate hungry. To find out the factors that influence the reception of poor rice, it is necessary to do a classification analysis of the subsidy food. The factors used in this study
include the business field of the head of the household (X_1); Employment Status (X_2); Homeownership status (X_3); Roof Building Materials (X_4); Main wall material of the house (X_5); Main Material of the house floor (X_6); Drinking Water Source (X_7) and Main Cooking Fuel (X_8). In this study, the classification of acceptance of poor rice is based on factors determined using the KNN method with a global weighting GINI Diversity Index.

2. Literature review

2.1. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is an algorithm with the principle of finding the closest distance between the data to be evaluated with its closest neighbor K in the training data [5].

Table 1. Classification Algorithms Using KNN

| Input: Training dataset, the observation of x and nearest neighbor the value of K; | Output: Prediction of the class label; |
|---|---|
| (1) Calculate the distance between training data and testing data | (2) Sort the distance from the smallest to the largest |
| (3) Get K nearest neighbor of x | (4) Define the majority of classes from the closest K K as the data testing class. |

2.2. Simple Matching Coefficient (SMC)

According to Chen and Guo (2015) [2], in applying k-NN to classify categorical data, a special way is needed to calculate the distance between existing data. One specific way to calculate the distance between data that is most often applied is to use the Simple Matching Coefficient (SMC). In this study, a classification method was used for categorical data using the nearest neighbor classification method by weighting the attributes using the Weighted Simple Matching Coefficient (WSMC) distance.

Suppose there is training data \(tr = \{z_1, z_2, ..., z_n\}\) with \(z_i = (x_i, y_i)\) for \(i = 1,2, ..., N\), with \(N\) is the number of samples, while \(x_i = (x_{i1}, x_{i2}, ..., x_{iK})\) are independent variables with \(K\) are the number of categories. While \(y_i\) it is dependent variables (class) from \(x_i\), with \(y_i \in \{1,2, ..., M\}\), with \(M\) is the number of classes that contain training data sets. Distance calculation using the Simple Matching Coefficient is as follows [6]:

\[
SMC(x_i, x_j) = \sum_{d=1}^{K} I(x_{id} \neq x_{jd})
\]

with \(I(\cdot)\) is the indicator function \(I(true) = 1; I(false) = 0\)

2.3. Weighted Simple Matching Coefficient (WSMC) with a weighted method using global gini diversity index

Weighted Simple Matching Coefficient (WSMC) is a statistical calculation that is almost the same as Simple Matching Coefficient which makes comparisons for each attribute with other attributes, but in its implementation, each attribute will not be calculated equally, but by giving weight to each attribute so that it indirectly reduce the influence of data which will disrupt classification. This calculation method can be used effectively to classify categorical data [7]. The new distance measurement will be denoted by WSMC\(_{global}\) according to the global weighting method used. In the global method, the attribute is associated with a weight vector \(\omega = (\omega_1, ..., \omega_d, ..., \omega_K)\) and the distance of WSMC between the object data of \(x_i\) and \(x_j\) given equation as follow as (Chen dan Guo, 2015) [2]:

\[
WSMC_{global}(x_i, x_j, \omega) = \sum_{d=1}^{K} \omega_d^{(GC)} \times I(x_{id} \neq x_{jd})
\]
The weighted $\omega_{d}^{(GG)}$ can be defined as:

$$
\omega_{d}^{(GG)} = \exp \left( - \frac{M}{M-1} \left( \sum_{s_{x},s_{y}} p(s_{d}) \times GG(s_{d}) \right) \right)
$$

(3)

The average Global GINI index of the variable of $d$-th can be defined as:

$$
AGG(d) = \frac{M}{M-1} \left( \sum_{s_{x},s_{y}} p(s_{d}) \times GG(s_{d}) \right)
$$

So the Eq. (3) can be written as:

$$
\omega_{d}^{(GG)} = \exp \left[ -AGG(d) \right]
$$

$GG(s_{d})$ counted by:

$$
GG(s_{d}) = 1 - \frac{\left( \sum_{m=1}^{M} p(m|s_{d}) \right)^{2}}{\sum_{(x,y)\in tr} I(x_{d} = s_{d})}
$$

$p(m|s_{d})$ is a probability to calculate the level of contribution of each attribute for each class $p(m|s_{d})$ counted by [8]:

$$
p(m|s_{d}) = \frac{\sum_{(x,y)\in ca} I(x_{d} = s_{d})}{\sum_{(x,y)\in tr} I(x_{d} = s_{d})}
$$

the equation $p(s_{d})$ can be written as:

$$
p(s_{d}) = \frac{1}{N} \sum_{(x,y)\in tr} I(x_{d} = s_{d})
$$

$N$ is the number of data. Classes for selected testing data are based:

$$
y = \arg \max_{m} \sum_{(s_{i},y)\in NN_{z}} I(m = y_{j})
$$

3. Method

The data used in this study is Semarang 2016 National Social Economic Survey data, independent variables used in this study are households receiving subsidy food (1 = Yes; 0 = No). Whereas the dependent variable used in this study is the business field of the head of the household ($X_{1}$) (0: not working; 1 = agriculture; 2 = mining; 3 = processing industry; 4: electricity and gas; 5: building; 6: trade, hotel and restaurant; 7: Transportation and warehousing; 8: Finance and insurance; 9: Services); Employment Status ($X_{2}$) (0: Not working; 1: self-employed; 2: trying to be assisted by temporary workers; 3: trying to be assisted by permanent workers; 4: Workers / employees / employees; 5: free workers; 6: family workers); Status of home ownership ($X_{3}$) (1: self-owned; 2: Contract / Rent; 3: Free rent; 4: Service; 5: Other); Roof Building Materials ($X_{4}$) (1: Concrete; 2: Ceramic Tiles; 3: Metal Tiles; 4: Clay Tiles; 5: Asbestos; 6: Zinc; 7: Bamboo; 8: Wood / Shingles; 9: Straw; 10 : Other); Main wall material of the house ($X_{5}$) (1: Wall; 2: wood; 3: log; 4: Other); Main Material of house floor ($X_{6}$) (1: Marble / Granite; 2: Ceramic; 3: Tile / Tile; 4: Cement / red brick; 5: Soil; 6: Other); Source of Drinking Water ($X_{7}$) (1: Branded bottled water; 2: Refill water; 3: Meter tapping; 4: Retail piping; 5: borehole / pump; 6: Protected well; 7: Unprotected well; 8: Eye protected water; 9: Other) and Main Cooking Fuel ($X_{8}$) (0: not cooking at home; 1: Electricity; 2: 5.5kg LPG; 3: 12kg LPG; 4: 3kg LPG; 5: kerosene; 6: firewood). The number of households used was 930 households with the category of receiving subsidy food totaling 227 households, while those not receiving subsidy food were 703 households. The number of training and testing data sharing is 80: 20 so that 744 is as training data and 186 is as testing data. In the training data, there are 182 households receiving subsidy food, while in testing data 45 households subsidize food.
4. Results and discussions
In the Nearest Neighbor k method by weighting the Global Gini Diversity Index, there are two stages carried out. The first stage is the stage of calculating the weight of each variable. The second stage is the data classification stage, the stage where the k-NN classification process is carried out by weighting the variables that have been obtained at the weight calculation stage. Calculation of the distance between data in this method uses Weighted Simple Matching Coefficient (WSMC) as follow as:

4.1. Steps of calculation the weight
Weight calculation for each variable is done by forming a weight vector \( \mathbf{\omega} = (\omega_1, \ldots, \omega_d, \ldots, \omega_k) \). Global weights are based on this index for attributes \( d \)-th can be donated as \( \omega_d^{(GG)} \). Here are the results of the calculation of the average index of the \( d \)-th variable or symbolized as \( AGG(d) \) as follow as:

| d-th variable | \( AGG(d) \) |
|---------------|---------------|
| 1             | 0.722889      |
| 2             | 0.626379      |
| 3             | 0.425709      |
| 4             | 0.626254      |
| 5             | 0.376815      |
| 6             | 0.411067      |
| 7             | 0.7303        |
| 8             | 0.415198      |

As follow as the weighted for each variable:
- 1-st variable
  \[ \omega_1^{(GG)} = e^{-AGG(1)} = e^{-0.3245} = 0.722889 \]
- 2-nd variable
  \[ \omega_2^{(GG)} = e^{-AGG(2)} = e^{-0.4678} = 0.626379 \]
- 3-rd variable
  \[ \omega_3^{(GG)} = e^{-AGG(3)} = e^{-0.854} = 0.425709 \]
- 4-th variable
  \[ \omega_4^{(GG)} = e^{-AGG(4)} = e^{-0.468} = 0.626254 \]
- 5-th variable
  \[ \omega_5^{(GG)} = e^{-AGG(5)} = e^{-0.976} = 0.376815 \]
- 6-th variable
  \[ \omega_6^{(GG)} = e^{-AGG(6)} = e^{-0.889} = 0.411067 \]
- 7-th variable
  \[ \omega_7^{(GG)} = e^{-AGG(7)} = e^{-0.3143} = 0.7303 \]
- 8-th variable
  \[ \omega_8^{(GG)} = e^{-AGG(8)} = e^{-0.879} = 0.415198 \]

So that the weighted as follows as
Based on the weighting, the variable that has the greatest influence on classification is X7, then continued X1, X4 and so on based on the largest weighting value.

4.2. The steps of Classification the data.

At the data classification stage, the classification process is carried out using the k-NN rule. Calculation of the distance between training data with testing data using WSMC distance. Calculations using the WSMC distance involve the indicator function of each testing data against each training data and the weight of the attributes that have been calculated at the weight calculation stage.

Following are the testing data used in calculating WSMC distances:

\[
\begin{align*}
&z_1 = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ 3 \end{pmatrix}, \\
&z_2 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 3 \end{pmatrix}, \\
&z_3 = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ 3 \end{pmatrix}, \\
&\ldots, \\
&z_{186} = \begin{pmatrix} 3 \\ 2 \\ \vdots \\ 4 \end{pmatrix}.
\end{align*}
\]

While the following are training data used in calculating WSMC distances:

\[
\begin{align*}
&z_1 = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ 3 \end{pmatrix}, \\
&z_2 = \begin{pmatrix} 8 \\ 1 \\ \vdots \\ 3 \end{pmatrix}, \\
&z_3 = \begin{pmatrix} 6 \\ 1 \\ \vdots \\ 3 \end{pmatrix}, \\
&\ldots, \\
&z_{744} = \begin{pmatrix} 3 \\ 2 \\ \vdots \\ 4 \end{pmatrix}.
\end{align*}
\]

The following formula is used in calculating WSMC distances:

\[
WSM_{global}(x_i, x_j, \omega) = \sum_{d=1}^{D} \omega_d \times I(x_{id} \neq x_{jd})
\]

with \(x_i\) is training data, \(x_j\) is testing data, and \(\omega\) is weighted of a variable. While \(I\) is an indicator function where if \(x_{id} \neq x_{jd}\) then the value of \(I\) is 1 and the value is 0 for \(x_{id} = x_{jd}\). So we get the WSMC distance calculation as follows:

WSMC distance between the 1-st testing data with the 1-th training data as follows as:

\[
WSM_{global}(x_1, x_1, \omega) = \sum_{d=1}^{8} \omega_d \times I(x_{1d} \neq x_{1d}) = [(0.722889 \times 0) + (0.626379 \times 1) + \cdots + (0.415198 \times 1)] = 4.08766
\]

WSMC distance between the 1-st testing data with the 2-nd training data as follows as:

\[
WSM_{global}(x_2, x_1, \omega) = \sum_{d=1}^{8} \omega_d \times I(x_{2d} \neq x_{1d}) = [(0.722889 \times 0) + (0.626379 \times 1) + \cdots + (0.415198 \times 1)] = 4.08766
\]

the calculation until 744-th training data and the 186-th testing data until 77-th training data.

WSMC distance between the 1-st testing data with the 744-th training data as follows as:

\[
WSM_{global}(x_{744}, x_1, \omega) = \sum_{d=1}^{8} \omega_d \times I(x_{744} \neq x_{1d}) = [(0.722889 \times 0) + (0.626379 \times 1) + \cdots + (0.415198 \times 1)] = 2.02345
\]
The results of the calculation of the first WSMC data testing distance to all training data that have been done sorting from the smallest to the largest value. Then the k value is selected by conducting a trial to obtain the Apparent Error Rate (APER) results as in Table 3. Based on Table 3, the K value that produces the smallest APER is K = 5.

| K  | 80: 20         |
|----|---------------|
| 1  | 0.1785        |
| 3  | 0.1800        |
| 5  | 0.1500        |
| 7  | 0.1675        |
| 9  | 0.1876        |
| 11 | 0.1875        |

Based on Table 3 it can be seen that for various K values, it turns out that the most optimal is for K = 5. As for various K values, the APER value was found to be greater than 0.15.

| Observation (Actual Class) | Predicted Class | Global Gini Diversity Index Weighted |
|----------------------------|-----------------|--------------------------------------|
| Kelas 0 (No Accepted)      | 125             | 16                                   |
| Kelas 1 (Accepted)         | 11              | 34                                   |

Table 4 shows the confusion matrix for the value of K = 5, with a comparison of training and testing is 80: 20. Based on the table the accuracy value is 85%, while the error rate is 15%.

5. Conclusion
Based on the results and discussions that have been carried out, the accuracy rate of receiving poor rice using the KNN method with a global weighting GINI diversity Index, obtained an accuracy value of 85%, with a prediction error rate of 15%. In this research paper, the accuracy of the classic KNN method needs to be compared.

References
[1] Gorunescu F 2011 Data Mining Concepts, Model, and Techniques. Jerman : Springer
[2] Chen L, Guo G 2015 Expert Systems with Applications Vol. 42 No. 6 Hal.3142–3149
[3] Badan Pusat Statistik 2010 Perhitungan Dan Analisis Kemiskinan Makro Indonesia Tahun 2010. Jakarta: Badan Pusat Statistik
[4] Badan Pusat Statistik 2014 Indikator kemiskinan. Jakarta: Badan Pusat Statistik
[5] Han J, Kamber M 2006 Data Mining : Concepts and Techniques. San Francisco Morgan Kaufmann Publishers
[6] Larose D T 2005 Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey A John Willey and Sons, Inc.
[7] Gou J, Ma H, Ou W, Zeng S, Rao Y & Yang H 2019 Expert Systems With Applications Vol. 115 Hal. 356-372