Implementation of Fuzzy K-Nearest Neighbor Method in Decision Support System for Identification of Under-five Children Nutritional Status Based on Anthropometry Index

Rahmalia Nurida[1], Endang Sugiharti[1], Alamsyah[1]
[1] Department of Computer Science, Faculty of Mathematics and Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: rahmalia.nurida7@gmail.com,

ARTICLE INFO

Article history
Received 17 June 2019
Revised 18 September 2019
Accepted 7 October 2019

Keywords
Fuzzy k-nearest neighbor
Decision support system
Anthropometry index
Nutritional Status

ABSTRACT

Nutritional status is one of the important factors in assessing the level of health and growth of infants and under-five children. But the present, there are still many problems caused by an imbalance in nutritional intake with the nutritional needs of children. K-Nearest Neighbor method in the previous studies showed the existence of prediction results with the problem. This study used a standard anthropometric index or body size to carry out the process of calculating nutritional status using the Fuzzy K-Nearest Neighbor method. Fuzzy is applied to reduce the problem in classification. Predictions produced are three categories of nutritional status, namely BB/U (weight according to age), TB/U (height according to age), and BB/TB (weight according to height). The k value taken for the classification process is k=10. Fuzzy K-Nearest Neighbor process has done by taking the closest Euclidean distance to the number k from the training data to the test data. The prediction class results of 95 out of 96 data are stated accordingly after the value of the membership is calculated. The accuracy of the test performed produces a cumulative accuracy of 98.96%. This study can be used as a reference for further research by adding training data with more complete class variations in each category of nutritional status to obtain more optimal accuracy.

1. Introduction

A decision support system is an 'interactive computer-based system' which serves to help decision making in the use of data and models for solving unstructured problems (Juliana, Jasmir and Jusia, 2017). That method is applied in decision support systems with the Fuzzy method. Fuzzy application is a method of searching linguistically data (Sugiharti, Arifudin & Putra, 2018). Fuzzy K-Nearest Neighbor method is one of the classification methods by combining fuzzy methods with the K-Nearest Neighbor algorithm. The Fuzzy K-Nearest Neighbor method was introduced by Keller et al. in 1985 by developing K-NN which was combined with fuzzy theory in conveying the definition of class labeling on predicted test data (Puspita, 2014). This method plays an important role in removing problems in classification. Calculation of the Fuzzy K-Nearest Neighbor method by calculating the distance between two data, that is adjusted data type. Where each data type has a formula (Prasetyo, 2012).

That is important to improve the quality of human resources today is an effort to prepare the generation with nutrition and health development ranging from fostering prospective mothers, caring for fetuses, infants, under-five children, and school children. This is meant by the early and continuous nutrition and health development, as well as the stimulation, carried out so that the formation of a quality generation is increasingly realized (Hadju, Metusalach & Karyadi, 1998). Good nutritional status plays a role in determining the success or failure of efforts to increase human resources (Putri & Sulastri, 2015). In Indonesia, the spectrum of malnutrition is very wide and occurs in all stages of life, among others in the form of Protein Energy Deficiency (KEP), micronutrient deficiencies, low weight babies, and growth disturbances as seen from indicators of height according to age. In terms of nutritional intake, disturbance growth indicates the cumulative effect of lack or insufficient energy intake, long-term macronutrients or micronutrients or the results of chronic infections or infections that occur repeatedly (Solihin, Anwar & Sukandar, 2013).

In fact, in the environment around us today there are still many babies and under-five children who experience nutritional problems. The health problems are generally caused by the consumption of unbalanced nutritional needs characterized by the disruption of physical and psychological children. There are various ways to assess nutritional status, one of which is the measurement of the human body known as anthropometry. Some types of anthropometry that have been used include bodyweight, height/body length, upper arm circumference, head circumference, chest circumference, and subcutaneous fat layer. In Indonesia, the commonly used anthropometry is body weight and height (Rismawan, Irawan, Prabowo & Kusumadewi, 2008). The assessment of nutritional status was based on z-score values calculated according to the anthropometric standard deviation of each under-five children. Nutritional status is calculated manually by looking at the categories of thresholds based on the index produced. This study applies the computer science field of decision support systems using the Fuzzy K-Nearest Neighbor method to facilitate the labeling of the results of under-five nutritional status by following the standard anthropometric index applicable in Indonesia.

2. Methods

2.1 Fuzzy-K Nearest Neighbor

The K-Nearest Neighbor algorithm is based on an analogy that compares test data and training data (Hidayah, Akhlis & Sugiharti, 2017). The KNN method is one of the simplest non-parametric classification methods, where a prediction class is determined based on the class most produced (Chen et al., 2013). When inputting test data, KNN will test the data based on known training data. The Euclidean formula is used to determine the proximity of the point or distance between data in the KNN (Sutarti, Putra & Sugiharti, 2019). Similar to the fuzzy theory, data has a membership function where the function shows the mapping of input data points into the value of membership degrees in intervals 0 and 1 (Hikmawati, Arifudin & Alamsyah, 2017). This method plays an important role in removing ambiguity in classification because it can represent the linguistic truth of a member of the set (Alamsyah & Muna, 2016). In its application, the Fuzzy K-Nearest Neighbor algorithm has several stages in classifying the data as follows.

Step 1

Determine the parameter k (number of closest neighbors) to calculate the FKNN method. A good k value can be chosen by parameter optimization, for example by using cross-validation. Special cases where classification is predicted based on the closest learning data (in other words, k=1) is called the nearest neighbor algorithm (D. Larose & C. Larose, 2015).

Step 2

Calculates the square of the Euclidean distance of the object against the data provided. The calculation of the distance test data to the training data used the Euclidean distance formula. Formula (1) is used to search for Euclidean distance.

$$d_1 = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \tag{1}$$

where:

x_1 : training data
x_2 : testing data
i : variable
d : distance
p : dimensions

Step 3

Sort the results object from the previous step into a group that has the smallest Euclidean distance. Data is sorted from the smallest Euclidean distance to the largest (ascending) to produce a category in the next step.

Step 4

Collect the Y category (the classification of the nearest neighbor value is based on the value of k). In the classification phase, the Euclidean distance order of the test data object against the training data is categorized based on the k value specified in the first step. Several data from the Euclidean distance were taken, this data being the Y category or classified as nearest neighbor.

Step 5

In the class prediction generated in the previous step, each class has calculated the value of the degree of membership. The class with the largest degree of membership will be the result of a prediction of testing. The calculation of membership values in each class uses the following function (Billyan, Bhawiyuga & Primananda, 2017).

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

where :

$u_i(x)$: the value of membership data x to class u_i

k : number of the closest neighbor

u_{ij} : membership value of neighboring data in k neighbor in class u_i , value 1 if the data training x_k belongs to the class u_i or 0 if it does not belong to the class u_i

$x - x_j$: distance from data x to data x_j in k nearest neighbor

m : weight exponent which is: m > 1

The flow chart of the Fuzzy K-Nearest Neighbor method is shown in Figure 1.

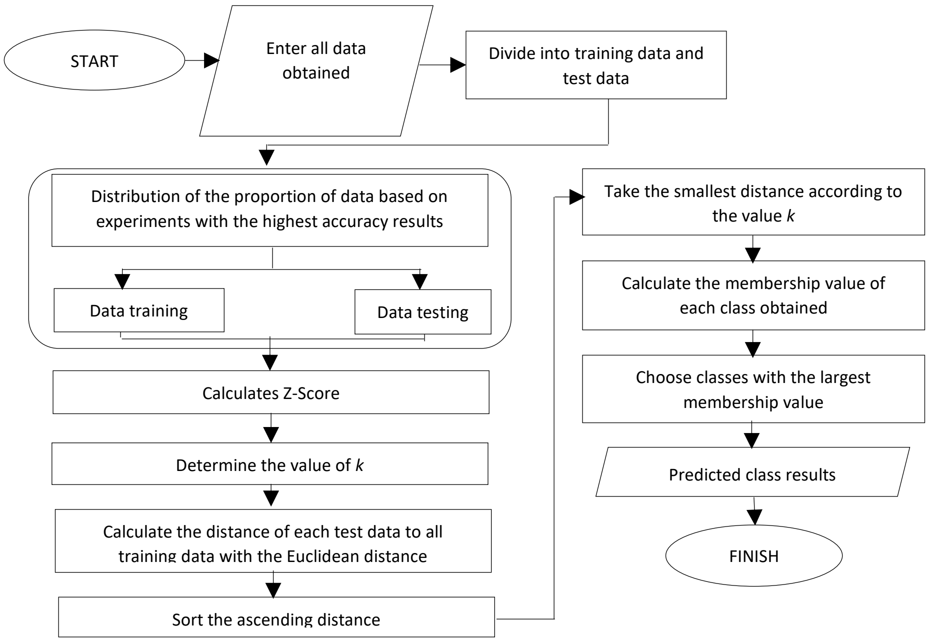

Figure 1. The Flowchart of the Fuzzy K-Nearest Neighbor Method

2.2 Calculation Accuracy

The method used to determine the final accuracy of tests performed is the confusion matrix for the multi-class method. This method is used to perform system calculations with many prediction classes (Manliguez, 2016). The difference from the multi-class confusion matrix with the ordinary confusion matrix is that the final results are calculated cumulative accuracy of the overall accuracy of all test data. Parameters of the accuracy are presented in Table 1.

Table 1. Confusion matrix

Classification	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TP)

where:

True Negative (TN) : if the prediction and actual results are negative

False Negative (FN) : if the positive prediction results, and the actual results negative

False Positive (FP) : if the negative prediction results, and the actual results positive

True Positive (TP) : if the predictive and actual results are positive.

Calculation of the total accuracy of the tests performed using the following formula.

$$Total\ Accuracy = \frac{TP+TN}{Number\ of\ data\ testing} \times 100\% \tag{3}$$

3. Results and Discussion

This study used standard anthropometric index under-five children published in the Decree of the Ministry of Health, Number: 1995/MENKES/SK/XII/2010. It is used as the basis for the Z-Score calculation before determining the nutritional status of under-five children.

3.1 The Results of the Data Testing

3.1.1 Data Partition

This data distribution manually by dividing the data into two parts namely is training data and test data. The training data in this study are used as experimental data while the test data is used as test data. The proportion of data used is training data as much as 60% (144 data) and test data as much as 40% (96 data). The percentage of data sharing used was taken based on experiments conducted five times. The proportion of final data taken based on calculations with the highest accuracy results.

3.1.2 Z-Score Standardization

Z-Score is used to take samples in data to determine the standard deviation value. To calculate the Z-Score, it needs to know beforehand the average, variance and standard deviation. As already stated that the standard used in determining the Z-Score in the study uses the anthropometric index of the Ministry of Health Decree*:* 1995/MENKES/SK/XII/2010. The following function is used for calculating Z-Score in determining the nutritional status of under-five children.

$$Z - Score = \frac{Observed\ Value - Median\ Value\ of\ the\ Reference\ Population}{Standard\ Deviation\ Value\ of\ the\ Reference\ Population} \tag{4}$$

3.1.3 Fuzzy K-Nearest Neighbor Process

The first step in calculating the Fuzzy K-Nearest Neighbor method determines the value of *k*, which is the value of the closest neighbor that will be used in the calculation. This study used the value of *k*= 10. The next step is to calculate the Euclidean distance between the test data and training data. After all the attributes in the training data and test data have been calculated Z-Score, then calculates the Euclidean distance from the test data against the Z-Score value of the overall training data. The results of the Euclidean distance calculation from the first test data to the overall training data are presented in Table 2.

Table 2. The Euclidean distance of the first test data on the overall training data

Code	Z-Score			Euclidean Distance		
	BB/U	TB/U	BB/TB	BB/U	TB/U	BB/TB
001	-2.25	-2.35	-1.67	1.083	0.304	2.667
002	-0.90	-3.00	1.17	0.267	0.952	0.167
003	0.00	0.91	-0.57	1.167	2.957	1.571
004	-2.45	-3.24	-1.20	1.288	1.192	2.200
005	0.00	-0.87	0.63	1.167	1.178	0.375
006	-0.39	-0.60	0.00	0.778	1.448	1.000
007	3.00	2.48	2.67	4.167	4.528	1.667
008	-1.86	-2.24	-1.13	0.690	0.195	2.125
:	:	:	:	:	:	:
143	-2.21	-1.97	-1.88	1.048	0.077	2.875
144	-1.17	-1.10	-0.25	0.000	0.952	1.250

The results obtained are then sorted according to the results of the smallest Euclidean distance to the largest Euclidean distance. Next is to take the minimum value according to the number of *k* values to determine the class category of the data. In this study, the value of *k*=10 was used, 10 data were taken with the smallest Euclidean distance value from each prediction class. In the weight according to age class, the 10 closest Euclidean distance values are presented in Table 3.

Table 3. The 10 Closest distance of weight class according to age from the first test data on training data

Code	Euclidean Distance	Nutritional Status
067	0.048	Good Nutrition
082	0.091	Malnutrition
055	0.100	Good Nutrition
019	0.130	Malnutrition
094	0.133	Good Nutrition
008	0.143	Good Nutrition
112	0.143	Malnutrition
114	0.182	Good Nutrition
071	0.182	Good Nutrition
076	0.200	Malnutrition

In this nutritional status identification system, three classes are produced. Namely the nutritional status class according to age, height according to age and weight according to height. In the last steps of the Fuzzy K-Nearest Neighbor method, we used formula (2) to calculate the degree of membership of each prediction class produced. From the calculation of each class according to the number of *k* values, the class with the highest value of membership is the prediction class of the test data. The prediction of nutritional status from 144 training data and 96 test data is shown in Table 4.

Table 4. Predicted Nutritional Status

Code	Gender	Predicted Nutritional Status		
		BB/U	TB/U	BB/TB
145	F	Good Nutrition	Stunted	Normal
146	M	Good Nutrition	Normal	Normal
147	F	Good Nutrition	Normal	Normal
148	M	Good Nutrition	Normal	Normal
149	M	Good Nutrition	Tall	Normal
150	M	Good Nutrition	Normal	Normal
151	M	Good Nutrition	Normal	Normal
152	M	Good Nutrition	Normal	Normal
:	:	:	:	:
239	F	Good Nutrition	Normal	Normal
240	M	Good Nutrition	Normal	Normal

3.2 Discussion

The data used in this study is in the form of anthropometric data of under-five children obtained from seven Maternal and Child Health Services in Krandon and Kaligangsa Village, Tegal City. The number of under-five children data was taken as 240 data from the Maternal and Child Health Services implementation from July to September 2018.

The first step is to determine the value of *k* or the value of the nearest neighbor. The value of *k* used in this study is k = 10. In this study, the data obtained was divided into 60% (144 data) as training data and 40% (96 data) as test data. Next is calculating the distance from the Z-Score value of each test data to the overall training data using the Euclidean Distance equation. The results of the entire Euclidean distance of the test data on the training data are then sorted from the smallest distance to the largest distance. It takes the 10 closest Euclidean distances; from the 10 closest distances you will get prediction classes from the test data. From each prediction class then the value of membership is calculated. The calculation of the value of membership is done to reduce the ambiguity of the prediction class produced. The class has the highest membership value and the result of the test data prediction class. The final step in this research is to calculate the accuracy of the tests that have been carried out. Accuracy calculation is done by using the confusion matrix method. The confusion matrix will give the results of the classification system performance based on the object correctly or incorrectly.

Table 5. FK-NN Method Accuracy Test Results use 60% Training Data and 40% Test Data

Calculations	Results
Number of Data Training	144
Number of Data Testing	96
Number of Correct Data	95
Number of Incorrect Data	1
Cumulative Accuracy Results	98.96%
Error Results	1.04%

Calculation of accuracy in this study using a confusion matrix that will provide decisions obtained from training data and test data. In this study, the multi-class confusion matrix was used because the prediction results consisted of three categories. The calculation of accuracy in this study was carried out by dividing the entire data into training data and test data. Prediction results from each test data are matched with the actual results to obtain the percentage accuracy of the tests. From the tests carried out, two test data were detected producing results that did not match the actual data. Test results and cumulative accuracy results using 60% training data and 40% test data can be seen in Table 5.

4. Conclusion

The use of Fuzzy algorithms combined with the K-Nearest Neighbor method can eliminate problems in classification. So that if the final result of the decision has the same ratio of the number of prediction classes, then the calculation of membership value serves to determine the prediction class is chosen. Decision making to identify nutritional status has three categories, namely BB/U (weight according to age), TB/U (height according to age), and BB/TB (weight according to height). In this study, 60% of the data were used as a training dataset and the rest was used as a testing dataset. From the tests carried out, the final results of the prediction class indicate 95 of the 96 test data showed predictive class results that were by the actual nutritional status class. So that the cumulative accuracy generated in this study is 98.96%.

References

Alamsyah, & Muna, I. H. (2016). Fuzzy Inference system method for performance evaluation of library staff and librarians. *Scientific Journal of Informatics*, 3(1), 88-98. doi:10.15294/sji.v3i1.6136

Billyan B. F., Bhawiyuga A., & Primananda R. (2017). Implementation of Fuzzy K-Nearest Neighbor classification for fingerprint access points in indoor positioning. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 1(11), 1195-1205.

Chen, H., Huang, C., Yu, X., Xu, X., Sun, X., Wang, G., & Wang, S. (2013). An efficient diagnosis system for detection for Parkinson's disease using K-Nearest Neighbor Approach. *ELSEVIER*, 40(1), 263-271. doi:10.1016/j.eswa.2012.07.014

Hadju V, Metusalach, & Karyadi D. (1998). Potential foods to increase physical growth thinking power and productivity to prevent degenerative diseases. *Widyakarya Nasional Pangan dan Gizi VI*, 17 – 20 February. LIPI, Jakarta. doi:10.25182/jgp.2006.1.2.8-16

Hidayah, M. R., Akhlis, I., & Sugiharti, E. (2017). Recognition number of The vehicle plate using the Otsu method and K-Nearest Neighbor Classification. *Scientific Journal of Informatics*, 4(1), 66-75. doi:10.15294/sji.v4i1.9503

Hikmawati, Z. S., Arifudin, R., & Alamsyah. (2017). Prediction the number of dengue hemorrhagic fever patients using the Fuzzy Tsukamoto method at the public health service of Purbalingga. *Scientific Journal of Informatics*, 4(2), 115-124. doi: 10.15294/sji.v4i2.10342

Juliana, Jasmir, & Jusia, P. A. (2017). Decision support system supplier selection using Analytical Hierarchy Process (AHP) Method. *Scientific Journal of Informatics,* 4(2), 158-168. doi: 10.15294/sji.v4i2.12015

Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics.* Canada: John Wiley & Sons, Inc.

Manliguez, C. (2016). Generalized confusion matrix for multiple classes. *ResearchGate*. University of the Philippines.

Prasetyo, E. (2012). Data mining concepts and applications using MatLab. Yogyakarta: Penerbit Andi.

Puspita, D. S. (2014). Prosperous family level classification using ordinal logistics and fuzzy k-nearest neighbor regression methods. *Jurnal Gaussian*, 3(4), 645-653. Retrieved from https://ejournal3.undip.ac.id/index.php/gaussian/article/view/8075

Putri, R. F., Sulastri D., (2015). Factors related to the nutritional status of toddler children in the work area of the Nanggalo Padang Health Center. *Jurnal Kesehatan Andalas,* 4(1), 254-261. doi:10.25077/jka.v4i1.231

Rismawan, T., Irawan, A. W., Prabowo, W., & Kusumadewi, S. (2008). Pocket PC based decision support system as a nutritional Status determinant using the KNN (K-Nearest Neighbor) Method. *Teknoin*, 13(2), 18-23. doi: 10.20885/.v13i2.793

Solihin, R. D. M., Anwar F., & Sukandar D. (2013). The relationship between nutritional status, cognitive development, and motor development in Preschool Children. *Nutrition and Food Research,* 36(1), 62-72.

Sugiharti, E., Arifudin, R., & Putra A.T. (2018). K-Means and Fuzzy Tahani as base of cattle data collection from manual card system to online information system. *Journal of Theorical and Applied Information Technology*, 96(21), 7176-7186.

Sutarti, S., Putra, A.T., & Sugiharti, E. (2019). Comparison of PCA and 2DPCA accuracy with K-Nearest Neighbor Classification in face image recognition. *Scientific Journal of Informatics*, 6(1), 64-72. doi: 10.15294/sji.v6i1.18553