Predicted Missing Imputation on Dengue Fever Spread Data with K-Nearest Neighbor (K-NN)

Taufiq Rizaldi[1], Fendik Eko Purnomo[2]

[1,2]Department of Information Technology, Politeknik Negeri Jember

taufiq_r@polije.ac.id [1], fendik_eko@polije.ac.id [2]

Abstract. Dengue Hemorrhagic Fever (DHF) is a disease caused by dengue virus with Aedes Aegypti intermediate. Based on a survey from the Health Office of Jember recorded during January 2015 out of 300 cases of DHF patients, 7 of them are died, that why the prediction of DHF distribution is needed for prevention of spreading. Parameters that used to determine the potential for the spread of DHF diseases are rainfall , rainy day, larva free and house index. However, the survey data is often incomplete, the missing imputation data resulted the process to predict the potential for the spread of DHF is still constrained. By used of K-Nearest Neighbor (K-NN) methods that can be used to predict the missing imputation data and complete it. Using the correlations between attributes attained on Euclidean Distance that shows better performance in terms of imputation accuracy. The method show MSE below 1 and MAPE around 10 – 16%.

Keywords : Dengue Hemorrhagic Fever, K-Nearest Neighbor, Missing Imputation

1. Introduction

Dengue Hemorrhagic Fever (DHF) is, DHF is caused by dengue virus with an intermediary of Aedes Aegypti mosquito [1]. The dengue virus itself consists of 4 types, namely DEN 1, DEN 2, DEN 3 and DEN 4 where each virus can cause mild dengue to fatal levels. The results of a survey conducted by the Jember District Health Office noted that in January 2015 out of 300 cases of DHF patients, 7 of them died, and in the first week of February 2015 there had been 42 cases. This shows the spread of DHF outbreaks can occur very quickly. One of the efforts made to prevent DHF is by predicting and mapping the possibility of DHF occurrence in each sub-district, the parameters used for the prediction process for DHF distribution include rainfall (CH), rainy day (HH), number of larva-free (ABJ) and house index (HI). However, in fact the survey data or data in the field often contains missing input data so that the data to be processed is incomplete. The thing that triggers the loss of value in the dataset or the lack of value in the data for certain attributes is the absence of response to the unit or item, this is a problem that occurs in some surveys and makes the prediction results of DHF distribution in each sub-district less appropriate.

From the data obtained, it can be seen that the missing imputation on the factor of spread of DHF in Jember regency can reach 10.5% in the period 2009-2012. One simple solution to overcome these problems is to remove the factor with the missing attribute value and apply the remaining dataset as input for the process subsequent analysis. However, this solution cannot be applied to the missing imputation case for DHF distribution data because it can cause biased results in the analysis process. The prediction data on DHF distribution will be useless if you don't use the appropriate method to get

the missing value. Generally missing data imputation is a process for applying a method that can generate a value to replace the missing value where the new value approaches the missing value. This study focuses on the application of methods that can be used to obtain estimated values of lost data.

2. Related Work

Research to apply a method that can be used to solve the problem of loss of data in a dataset that has been done a lot. Monte carlo method to predict the spread of DHF using the Monte-Carlo method to predict and solve the missing imputation problem, where the use of the method can reduce the Mean Square Error to 2.6% [2]. Implementing Interactive-KNN to overcome data loss problems in the Trash Pickup Logistics Management System (TPLMS) get the Rate of Reliable Imputation reaches 0.70[3]. From several methods that have been applied by several intelligent techniques such as k-nearest neighbors (KNN)[4], Bayesian principal-component analysis (BPCA) and local least squares (LLS) can be applied to find the missing value by estimating the value based on the value on existing data sets without using data with missing values. In this study the chosen method is KNN method because it is known to be more effective and simpler than other methods.

2.1. DBD Spatial Factors Affecting DHF

Some spatial factors that influence the spread of DHF and used in this study are[5][6] :

a. Rainfall Index (ICH) is the multiplication of rainfall and rainy days divided by the number of days in the month. ICH does not directly affect mosquito reproduction, but affects the ideal rainfall. Ideal rainfall means that rainwater does not cause flooding and stagnant water in a medium / media that is a safe and relatively clean place for mosquito breeding.
b. Rainy Day (HH) is the multiplication of rainfall and rainy day (HH) which is multiplied by the number of days in the month. The more HH in one month, the higher the potential for the spread of DHF (RI Ministry of Health, 2010).
c. The larva-free number (ABJ) is the number of larvae-free buildings compared to the number of buildings examined. If ABJ (> 95%) is expected that DHF transmission can be prevented or reduced.
d. House Index (HI) is one of the measures used to determine the spread of larvae of Aedes Aegypti mosquitoes where if the HI value is less than 5% then the possibility of spreading the larvae of Aedes Aegypti mosquitoes is still very small otherwise if the HI condition above 5% is an indicator of the level of susceptibility to transmission DHF. HI is obtained from the number of positive buildings, there is DHF divided by the number of buildings examined multiplied by 100%.

2.2. K Nearest Neighbor (KNN) Imputation

K Nearest Neighbor (KNN) is known as one of the classic methods that can be used to overcome the problem of loss of value in a dataset. Similarity values are used to determine a case in a group or closer to a particular group[7].The KNN method utilizes data groups that have complete values to predict the value of lost data. The KNN steps to get the missing data values are as follows [3][4].

a. Determining the value of the K parameter, K is the number of closest observations that will be used. The determination of the K value must be based on empirical reasons and based on the results of the experiment.
b. Every data with missing values was calculated the distance between observations containing missing data with complete observation on the j variable that does not contain missing data with other j variables that correspond to the eucludean distance formula, namely:

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{m} (x_{aj} - x_{bj})^2} \quad (1)$$

$d(x_a,x_b)$ is the distance between observations containing missing data and observations that do not contain missing data, x_{aj} is the value of the j variable for each observation that contains missing data with j = 1,2, …, m, x_{bj} is the value of other variables in each observation that does not contain missing data j = 1,2, …,m.

c. Sorting distances based on observations that have the greatest distance value until observations that have the smallest distance value.
d. Determine the closest observation K based on the smallest distance value.
e. Performing missing data imputation by calculating the weight mean estimation (WME) value on the closest observation K that does not contain missing data values with the formula:

$$x_j = \frac{\sum_{k=1}^{K} w_k v_k}{\sum_{k=1}^{K} w_k} \quad (2)$$

where is the x_j estimated weighted average, is the value v_k in the complete data on variables containing missing data based on observations of k, K is the number of closest observations used, k is the observation of K, w_k is the weight of observation of the nearest K-neighbor with formula $w_k = \frac{1}{d\,(x_{ak}x_{bk})^2}$, where d is the observation distance K.

f. Perform the process of missing data imputation on observations containing missing values with the average values obtained.

3. Methodlogy

a. Fundamental Dataset

The dataset used in this research is rainfall (CH), rainy day (HH), larva-free number (ABJ) and house index (HI) of Jember district every month in 2009 - 2012 with the percentage of lost data in the range of 10.56 % of the total data, for data testing the data that has been obtained is converted into a two-dimensional table as shown in table 1 where A = {A_1, A_2, A_3, A_4, ..., A_n} are attributes or variables of the research, B = {B_1, B_2, B_3, B_4, ..., B_n} are the data in the n-district for each attribute, X_{ij} is the value of the attribute An on the month to B_n, and M is the amount of data missing from attribute An in sub-district B_n.

Table 1. Dataset

B	A_1	A_2	A_3	A_4	M
B_1	266	12	100	0	0
B_2	201	9	100	0	0
B_3	166	7	100	0	0
B_4	239	8	?	0	1
B_5	220	6	95	5	0
B_6	330	14	100	0	0
B_7	276	22	100	0	0
B_8	334	13	100	0	0
B_9	246	17	100	0	0
B_{10}	244	16	100	0	0
B_{11}	256	15	100	0	0
B_{12}	305	16	100	0	0
B_{13}	199	13	100	0	0
B_{14}	134	18	100	0	0
B_{15}	220	13	100	0	0
B_{16}	302	16	100	0	0
B_{17}	207	22	100	0	0
B_{18}	195	15	?	?	2
B_{19}	190	17	100	0	0

B_{20}	344	19	100	0	0
B_{21}	327	22	100	0	0
B_{22}	280	18	100	0	0
B_{23}	360	17	100	0	0
B_{24}	267	19	100	0	0
B_{25}	310	20	100	0	0
B_{26}	648	14	100	0	0
B_{27}	456	17	100	0	0
B_{28}	458	15	100	0	0
B_{29}	247	17	98	6	0
B_{30}	140	20	?	?	2
B_{31}	311	19	?	?	2

4. Experiment and Result

In this section experimental results from the KNN method will be displayed on the dataset for DHF spreading factors. In the initial stages the dataset as shown in Table 1 is separated into 2 datasets, which are dataset with missing value and without missing value. Suppose the method will be used to find the missing value in the B_4 Attribute A_3 data, then euclidean n calculation is done for the data without missing value, then the calculation results are sorted as in Table 2.

Table 2. Euclidean Distance Sorting Results for B_4 Attribute A_3 data.

No	B	Euclidean
1	B_{10}	9.486832981
2	B_9	11.44552314
3	B_{29}	13.78404875
4	B_{11}	18.41195264
5	B_{15}	19.67231557
6	B_5	19.74841766
7	B_1	27.31300057
8	B_{24}	30.09983389
9	B_{17}	34.94281042
10	B_2	38.02630668
11	B_7	39.57271787
12	B_{13}	40.32369031
13	B_{22}	42.21374184
14	B_{19}	49.82971001
15	B_{16}	63.51377803
16	B_{12}	66.49060084
17	B_{25}	72.01388755
18	B_3	73.01369735
19	B_{21}	89.11228871
20	B_6	91.20307012
21	B_8	95.13674369
22	B_{14}	105.4798559
23	B_{20}	105.579354
24	B_{23}	121.3383699
25	B_{27}	217.1888579
26	B_{28}	219.1141255
27	B_{26}	409.0452298

To get a comparison of the results of imputation data used five K values, namely NN, 3, 5, 7 and 9 then missing data imputation by calculating the weight mean estimation (WME) value on the closest observation K. The results of imputation are shown in table 3.

Table 3. Imputation Results

K	WME
NN	100
k-NN 3	73.97090926
k-NN 5	88.21870703
k-NN 7	89.21498831
k-NN 9	97.50108726

To evaluate the calculation of MSE (Mean Square Error) with and MAPE (Mean Absolute Percentage Error) for each value of K. The smaller MSE value shows the better results while the MAPE value is considered very good if the MAPE value is <10%, while the results of good imputation if MAPE value between 10% and 20%. The results of the evaluation as shown in Table 4. From Table 4, the average MSE value is still below 1 and the average MAPE value is still below 10%, only when K = 3 the MAPE value is above 10%.

Table 5. Average MSE and MAPE values in missing data.

K	MSE	MAPE
NN	0.0745	10.7040816 %
3	0.2587	16.9346939 %
5	0.1151	10.7275863 %
7	0.1471	12.3040816 %
9	0.1057	13.1265306 %

5. Conclusion

The performance of the K-Nearest Neighbor (K-NN) method for imputation of lost data in a group of data has sufficient performance satisfying based on MSE (Mean Square Error) calculations where the smaller the MSE value the better, the average MSE value is still below 1 that is for NN = 0.0745, k-3 = 0.2587, k-5 = 0.1151, k-7 = 0.1471, and k-9 = 0.1057. Whereas for MAPE where if the MAPE value <10% means it is very good and if between 10% - 20% has a good meaning, get results for NN = 5.7040816%, k-3 = 11.9346939%, k-5 = 5.7275863%, k-7 = 7.3040816%, and k-9 = 8.1265306%.

References

[1] Bitari, I Putu Dody Lesmana, Slamet Yulianto, Model Potential Spread of Disease Fever Dengue in Jember Method Using Fuzzy, Prosiding Conference on Smart-Green Technology in Electrical and Information Systems. Bali, 14-15 November 2013.

[2] Roziqin,M. Choirur, Achmad Basuki, Tri Harsono. 2016. A Prediction System of Dengue Fever Using Monte Carlo Method. EMITTER International Journal of Engineering Technology Vol. 4, No. 1, June 2016. Surabaya.

[3] Zhu, Ming, Xingbing Cheng. 2015. Iterative KNN Imputation Based on GRA for Missing Values in TPLMS. 4th International Conference on Computer Science and Network Technology (ICCSNT 2015). Harbin, China.

[4] R. Jornsten, H. Y. Wang, and W. J. Welsh. DNA microarray data imputation and significance analysis of differential expression. Bioinformatics, 2005, 21:4155-4161.

[5] Ministry of Health. 2010. Demam Berdarah Dengue. Indonesia: Indonesian Ministry of Health.

[6] Directorate General of Disease Control and Environmental Health. 2011. Demam Berdarah. Jakarta.

[7] Rizaldi. T, M.A. Muslim, E. Yudaningtyas. 2014. Knowledge Management System untuk Diagnosis Infeksi Nosokomial. Jurnal EECCIS 8(2),105-110. Universitas Brawijaya:Malang.

Acknowledgments
Our thank goes to Department of Information Technology, Politeknik Negeri Jember, who has helped support for this research.