Implementation of *Modified K-Nearest Neighbor* for Diagnosis of Liver Patients

Alwis Nazir, Lia Anggraini, Elvianti, Suwanto Sanjaya, Fadhilla Syafria
Department of Informatics, Faculty of Science and Technology
State Islamic University of Sultan Syarif Kasim Riau
Pekanbaru, Indonesia
Email : alwis.nazir@uin-suska.ac.id

Abstract— Number of patients with liver disease in the world is very high. In the early stages, liver disease is difficult to detect. Early diagnosis of the liver disease may help in preventing and treating sufferers. To diagnose liver disease can be done with a blood test. Based on data from this analysis, the results can assist in determining patients with liver disease. This study uses data Indian Liver Patient Dataset (ILPD) taken from the UCI Machine Learning Repository. We used Modified k-Nearest Neighbor to classify into two classes, namely sufferers and non-sufferers. The amounts of data used in this study were 583 records. Tests performed by dividing the training data and test data to 50:50, 60:40, 70:30 and 80:20. Results of tests performed can classify with a good degree of accuracy reached 85.14% with a ratio of 70:30 and k = 3.

Keywords: ILPD, Liver, *Modified k-Nearest Neighbor*.

Patients with liver disease in the world in general is still relatively very high, as evidenced by the World Life Expectancy data shows liver disease ranks fourteenth to the death toll reached 1,020,891 in 2014 [1]. Liver disease classified as a disease that is difficult to recognize at an early stage, even when it has spread. Early diagnosis of liver disease is needed to be able to assist in preventing and treating sufferers of the disease [2].

With a large number of deaths from liver disease, then a diagnostic tool being developed to prevent and reduce mortality. In diagnosing liver disease need blood tests to analyze levels of the enzymes contained in the blood [3]. Based on the results of the blood test can be seen a patient suffering from liver disease or not. With these problems, naturally the technology as a diagnostic tool is needed to help people cope with liver disease early. To help diagnose a disease can use data mining method based on data obtained from the analysis ever undertaken.

Some studies related to the diagnosis of liver disease, namely diagnosing liver disease using a rule based classifier [4]. Rule-based algorithms used are ZeroR, OneR, RIPPER and C4.5. Rule-based algorithms can be used and implemented to create an automated system for detecting liver disease. The ZeroR algorithm has highest accuracy and precision value. C4.5 algorithms have the greatest value of sensitivity and specificity among other algorithms. In another study[3], classify liver disease with multiple categories of classification algorithms. The algorithm used are J48, Multilayer Perceptron (MLP), Random Forest (RF), Multiple Linear Regression (MLR), Genetic Programming and Support Vector Machine (SVM). The result is Random Forest method is the best relative model from other models with an accuracy of 89.11%.

In this study, we used a classification method to diagnose of liver disease that is Modified k-Nearest Neighbor (Mk-NN). In a study [5], Modified k-Nearest Neighbor is a modification of the k-Nearest Neighbor algorithm with some additional processes, namely the validity of the training data and voting weight. Parvin did a comparison level of accuracy between Modified k-Nearest Neighbor and k-Nearest Neighbor on multiple datasets. The results obtained Mk-NN level accuracy better than k-Nearest Neighbor (k-NN). Meanwhile, another study [7] in the classification of soybean plants disease showed an accuracy rate of 92.4% with a value of k = 3.

The main idea of Modified K-Nearest Neighbor (Mk-NN) is to classify test sample based on label that frequently appear on the label of neighbors. The level of accuracy of Modified k-Nearest Neighbor (Mk-NN) is better than k-Nearest Neighbor (k-NN) which only based on Euclidian distance [5].

Based on these issues, we conducted research on how Modified k-Nearest Neighbor can identify liver disease. With this method, it will classify someone into the category of patient with liver disease or not.

I. ANALYSIS

A. Data Preparation

a. We obtained Indian Liver Patient Dataset (ILPD) in 2012 from UCI *Machine Learning Repository*.

b. The amount of data has 583 records with 11 attributes. This data consists of 416 patients suffering from liver disease and 167 patients who did not suffer. The class classification defined by the numerical as 1 (patients with liver disease) and 2 (not patients).

Below the table of attributes which is used for the classification.

TABLE I. THE ATTRIBUTES OF LIVER DISEASE PATIENT CLASSIFICATION

No	Variable	Info	Type of Data
1	Age	Patient's Age	Integer
2	Gender	Patient's Gender	Binominal
3	TB	*Total Bilirubin*	Integer
4	DB	*Direct Bilirubin*	Integer
5	Alkphos	*Alkaline Phospotase*	Integer
6	Sgpt	*Alamine Aminotransferase*	Integer
7	Sgot	*Aspartate Aminotransferase*	Integer
8	TP	*Total Protiens*	Integer
9	ALB	*Albumin*	Integer
10	A/G	Rasio *Albumin dan Globulin*	Integer
11	*Selector*	Class Label	Binominal

Here's a sample of our original data which used for diagnose of liver patients.

TABLE II. SAMPLE OF ORIGINAL DATA

No	PA	PG	TB	DB	Alk	Sgpt	Sgot	TP	Alb	A/G	Class
1	65	2	0.7	0.1	187	16	18	6.8	3.3	0.9	1
2	62	1	10.9	5.5	699	64	100	7.5	3.2	0.74	1
3	62	1	7.3	4.1	490	60	68	7	3.3	0.89	1
4	58	1	1	0.4	182	14	20	6.8	3.4	1	1
5	72	1	3.9	2	195	27	59	7.3	2.4	0.4	1
6	46	1	1.8	0.7	208	19	14	7.6	4.4	1.3	1
7	26	2	0.9	0.2	154	16	12	7	3.5	1	1
8	29	2	0.9	0.3	202	14	11	6.7	3.6	1.1	1
9	17	1	0.9	0.3	202	22	19	7.4	4.1	1.2	2
10	55	1	0.7	0.2	290	53	58	6.8	3.4	1	1
…	…	…	…	…	…	…	…	…	…	…	…
583	38	1	1	0.3	216	21	24	7.3	4.4	1.5	2

B. *Data Mining Stage Analysis*

1) *Data Cleansing*

At this stage, it will perform data cleansing. The cleansing process is to transform the data which has some missing values or incomplete data into default values by using median value.

There are four missing values is found in the original data so the data cleansing needs to be done. The process is using average value ($\bar{X}$) from previous value (Xi) and next data (Xj) of the related attributes.

$$\bar{X}_{242} = \frac{x_i + x_j}{2} = \frac{1+1}{2} = 1$$

$$\bar{X}_{254} = \frac{x_i + x_j}{2} = \frac{1.2+1.4}{2} = \mathbf{1.3}$$

2) *Data Transformation*

The next process about the data should be normalized. The purpose of this normalization, data needs to be in the range [0-1] so that the distribution of the data is not too far.

$$v'_i = \frac{v_i - min_A}{(Max_A - Min_A)}(new\_max_A - new\_min_A) + new\_min_A$$

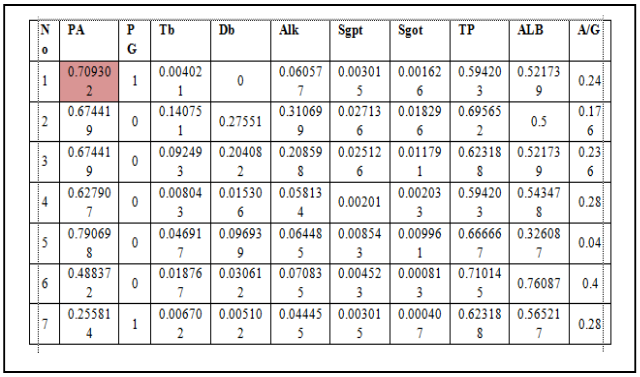
$$v'_i = \frac{65-4}{(90-4)}(1-0)+0$$

$$v'_i = \frac{61}{86}$$

$$= \mathbf{0.709302}$$

The result of data transformation:

TABLE III. DATA TRANSFORMATION



No	PA	PG	Tb	Db	Alk	Sgpt	Sgot	TP	ALB	A/G
1	0.709302	1	0.004021	0	0.060577	0.003015	0.001626	0.594203	0.521739	0.24
2	0.674419	0	0.140751	0.27551	0.310699	0.027136	0.018296	0.695652	0.5	0.176
3	0.674419	0	0.092493	0.204082	0.208598	0.025126	0.011791	0.623188	0.521739	0.236
4	0.627907	0	0.008043	0.015306	0.058134	0.00201	0.002033	0.594203	0.543478	0.28
5	0.790698	0	0.046917	0.096939	0.064485	0.008543	0.009961	0.666667	0.326087	0.04
6	0.488372	0	0.018767	0.030612	0.070835	0.004523	0.000813	0.710145	0.76087	0.4
7	0.255814	1	0.006702	0.005102	0.044455	0.003015	0.000407	0.623188	0.565217	0.28

3) *Classification using Modified k-Nearest Neighbor*

Based on the data attributes that have been obtained in the previous process, then this section will explain how to use Mk-NN in classification data. For more details on how Mk-NN algorithm works, described in the flowchart shown in figure below.

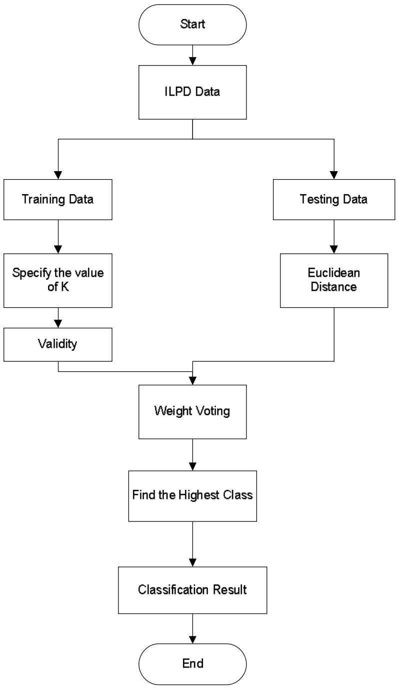


Fig. 1. Classification of Liver Disease with Mk-NN

The following description of classification of liver disease patients with Mk-NN algorithm :

1. *Data Divide*

All data is divided into training data and testing data. This data is divided by a ratio of 50:50.

TABLE IV. TRAINING DATA

No	PA	PG	tb	Db	Alk	Sgpt	Sgot	TP	ALB	A/G	Class
1	65	2	0.7	0.1	187	16	18	6.8	3.3	0.9	1
2	62	1	10.9	5.5	699	64	100	7.5	3.2	0.74	1
3	62	1	7.3	4.1	490	60	68	7	3.3	0.89	1
4	58	1	1	0.4	182	14	20	6.8	3.4	1	1
5	72	1	3.9	2	195	27	59	7.3	2.4	0.4	1
6	46	1	1.8	0.7	208	19	14	7.6	4.4	1.3	1
7	26	2	0.9	0.2	154	16	12	7	3.5	1	1
8	29	2	0.9	0.3	202	14	11	6.7	3.6	1.1	1
9	17	1	0.9	0.3	202	22	19	7.4	4.1	1.2	2
10	55	1	0.7	0.2	290	53	58	6.8	3.4	1	1
…	…	…	…	…	…	…	…	…	…	…	…
579	38	1	1	0.3	216	21	24	7.3	4.4	1.5	2

We took only four testing data as a sample for calculation of liver disease classification.

TABLE V. TESTING DATA

No	PA	PG	TB	DB	Alk	Sgpt	Sgot	TP	Alb	A/G	Class
1	38	1	1.8	0.8	342	168	441	7.6	4.4	1.3	1
2	40	1	1.1	0.3	230	1630	960	4.9	2.8	1.3	1
3	70	1	1.4	0.6	146	12	24	6.2	3.8	1.58	2
4	32	1	23	11.3	300	482	275	7.1	3.5	0.9	1

2. *Calculation of Validity*

This calculation starts with the determination of value of k. Then calculate each variable for each class in the training data. The following calculation to find the validity value for k = 3

$$\text{Validitas (x)} = \frac{1}{k}\sum_{i=1}^{k} S\,(lbl(x),(lbl(N_i(x))$$

$$\text{Data1} = \frac{1}{3}(1+0+1) = \mathbf{0.67}$$

$$\text{Data2} = \frac{1}{3}(1+1+1) = \mathbf{1}$$

Perform these steps for each training data. The table below is the results for the overall validity of the training data.

TABLE VI. TESTING DATA VALIDITY

Data Number	Validity
1	0.67
2	1.00
3	1.00
4	1.00
5	1.00
6	1.00
7	0.67
8	0.67
9	0.00
10	0.67
…	…
579	0.00

3. *Euclidean Distance*

Calculate the Euclidean distance of each parameter training data and testing data. Here's the formula for calculating the Euclidean distance (de),

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2}$$

$$\begin{aligned} \text{d(1,1)} &= \sqrt{(65-38)^2 + (2-1)^2 + (0{,}7-1{,}8)^2 + (0{,}1-0{,}8)^2 + (187-342)^2 + (16-168)^2 +} \\ &\quad \sqrt{(18-441)^2 + (6{,}8-7{,}6)^2 + (3{,}3-4{,}4)^2 + (0{,}9-1{,}3)^2} \\ &= \sqrt{729 + 1 + 1{,}21 + 0{,}49 + 24025 + 23104 + 178929 + 0.64 + 1{,}21 + 0.16} \\ &= \sqrt{226791{,}71} = \mathbf{476{,}227} \end{aligned}$$

Perform these steps for each of training data to testing data. The table below is the result of Euclidean distance (d_e).

TABLE VII. EUCLIDEAN DISTANCE RESULT

No	d_e testing data 1	d_e testing data 2	d_e testing data 3	d_e testing data 4
1	476.227	1869.449	41.962	545.607
2	505.202	1847.304	560.776	604.674
3	416.312	1824.474	350.269	508.165
4	476.403	1870.211	38.224	547.039
5	434.259	1839.477	62.191	516.487
6	471.752	1868.360	67.613	540.105

No	d_e testing data 1	d_e testing data 2	$d_e$testing data 3	d_e testing data 4
7	492.581	1873.413	46.502	555.238
8	477.808	1874.291	70.653	546.757
9	468.447	1863.455	77.923	536.260
10	403.621	1817.790	154.271	482.058
…	…	…	…	…
579	459.755	1861.501	77.504	532.184

4. *Weight Voting*

Calculate weight voting using validity and Euclidean distance value of each variable for each class in training data. The following calculation to find the value of voting weight:

$$\mathrm{W}(i) = Validity(i) * \frac{1}{d_e + 0{,}5}$$

$$\mathrm{W1{,}1(trainingdata1, testingdata1)} = 0.67 * \frac{1}{476.227 + 0{,}5}$$

$$= \mathbf{0.0041953}$$

$$\mathrm{W1{,}2(trainingdata1, testingdata2)} = 0.67 * \frac{1}{1869.449 + 0{,}5}$$

$$= 0.000358299$$

$$\mathrm{W1{,}3(trainingdata1, testingdata3)} = 0.67 * \frac{1}{41.962 + 0{,}5}$$

$$= 0.0157788$$

$$\mathrm{W1{,}4(trainingdata1, testingdata4)} = 0.67 * \frac{1}{545.607 + 0{,}5}$$

$$= 0.00122687$$

Perform these steps for each of training data to testing data. The table below is the result of weight voting.

TABLE VIII. WEIGHT VOTING RESULT

No	WV testing data 1	WV testing data 2	WV testing data 3	WV testing data 4
1	0.00140542	0.000358299	0.0157788	0.00122687
2	0.00197745	0.000541183	0.00178165	0.00165242
3	0.00239916	0.000547953	0.00285088	0.00196593
4	0.00209686	0.000534556	0.0258239	0.00182635
5	0.00230013	0.000543485	0.0159511	0.00193428
6	0.00211751	0.000535086	0.0146814	0.00184978
7	0.0013588	0.000357541	0.0142546	0.0012056
8	0.00140077	0.000357373	0.00941631	0.00122429
9	0	0	0	0
10	0.00165792	0.000368478	0.00432898	0.00138844
…	…	…	…	…
579	0	0	0	0

5. *Weight Voting Majority*

From the results of weight voting, do a search of the majority or dominant class based on k which has been used (validity formula). The results are the classification of patients with liver disease and non liver disease patient.

The table below is the result of the highest k of weight voting.

TABLE IX. WEIGHT VOTING RESULT

No	WV testing data 1	WV testing data 2	WV testing data 3	WV testing data 4
1	0.01033	0.0027	0.1133	0.0083
2	0.00766	0.0017	0.0731	0.0039
3	0.00714	0.0014	0.0726	0.0036

After the highest k of weight voting is obtained, then search a class from each data of the highest weight voting. Afterwards find a majority of each class of weight voting. The original class of weight voting and its majority can be seen in the following table.

TABLE X. THE ORIGINAL CLASS OF WEIGHT VOTING

No	WV testing data 1	WV testing data 2	WV testing data 3	WV testing data 4
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
Majority	**1**	**1**	**1**	**1**

Having obtained the majority of the class, then we did comparison between classification result and real class of testing data.

TABLE XI. CLASSIFICATION RESULT

No	Real Class	Class of Classification	Prediction
1	1	1	True
2	1	1	True
3	2	1	Wrong
4	1	1	True

II. RESULTS

We used confusion matrix as a testing method to calculate method accuracy which had been implemented for this study. Implementation of tests performed as follows:

A. Testing without Normalization

Figure 2 is a chart of testing results without normalization of the data partition with ratio 50:50, 60:40, 70:30 and 80:20 with using parameter k = 1 to k = 10.

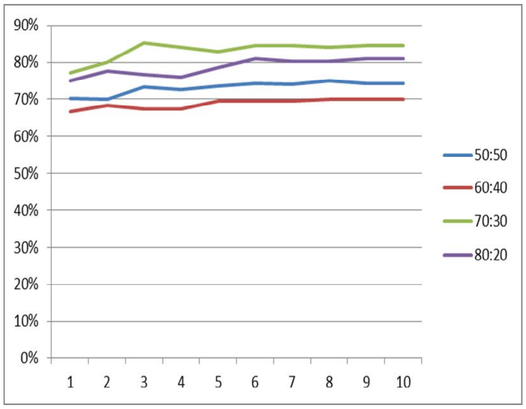


Fig. 2. Testing Results without Normalization

The highest accuracy of data partition with ratio 50:50 is 75% at k = 8 and the average accuracy is 73.14%. In 60:40 ratio, the highest accuracy data is 69.96% at k = 8, k = 9 and k = 10 and the average accuracy is 68.81%. In 70:30 ratio, the highest accuracy is 85.14% at k = 3 and the average accuracy is 83.14%. In 80:20 ratio, the highest accuracy data sharing is 81.03% at k = 6, k = 9 and k = 10 and the average accuracy is 78.70%.

Below the table of confusion matrix of classification with 50:50 ratio using parameter k=1.

TABLE XII. CONFUSION MATRIX WITHOUT NORMALIZATION

		Actual	
		Sufferers	Non Sufferers
Prediction	Sufferers	186	31
	Non Sufferers	56	19

Based on the confusion matrix table, accuracy values can be calculated as follows:

$$Accuracy = \frac{(186+19)}{(186+19+56+31)}$$

$$Accuracy = 70.2\%$$

B. *Testing with Normalization*

Figure 3 is a chart of testing results with normalization of the data partition with ratio 50:50, 60:40, 70:30 and 80:20 with using parameter k = 1 to k = 10.

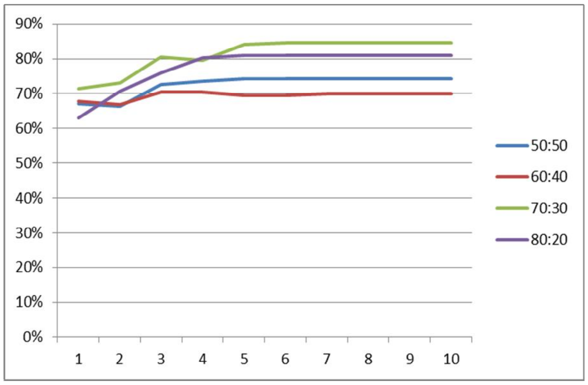


Fig. 3. Testing Results with Normalization

The highest accuracy of data partition with ratio 50:50 is 74,31% at k = 5 up to k = 10 and the average accuracy is 72,56%. In 60:40 ratio, the highest accuracy data is 70,39% at k = 3 and k = 4 and the average accuracy is 69,45%. In 70:30 ratio, the highest accuracy is 84,57% at k = 6 up to k = 10 and the average accuracy is 81.14%. In 80:20 ratio, the highest accuracy data sharing is 81.03% at k = 5 up to k =10 and the average accuracy is 77.58%.

Below the table of confusion matrix of classification with 50:50 ratio using parameter k=1.

TABLE XIII. CONFUSION MATRIX WITH NORMALIZATION

		Actual	
		Sufferers	Non Sufferers
Prediction	Sufferers	173	44
	Non Sufferers	52	23

Based on the confusion matrix table, accuracy values can be calculated as follows:

$$Accuracy = \frac{(186+23)}{(173+23+52+44)}\ x\ 100\ \%$$

$$Accuracy = 67{,}12\%$$

III. CONCLUSION

In this study, we proposed a classification which can classify liver disease sufferers and non-sufferers. From the result analysis we concluded, as follows:

1. The best accuracy using testing without normalization by value of the parameter k = 3 and using the percentage distribution of training data and test data 70:30 with a value of 85.14%.
2. In our tests on k= 5 up to k = 10 without normalization or normalization, accuracy results which we obtained are virtually identical on all data sharing.

For further study, it is recommended to use primary data so the classification is more varied and there will be a balance amount of class. In addition, the study should use other methods such as discriminant analysis and these two methods (Mk-NN and discriminant analysis) can be compared.

REFERENCES

[1] Expentacy, W. L. (2014). World Rangking Total Deaths. Retrieved July 29, 2016, from http://www.worllifeexpentacy.com/world-rangkings-total-deaths

[2] Satyatama, R. D., Indriati, & Setiawan, B. D. (2013). Incomplete Data Classification of Liver Disease Using Voting Featured Interval-5 Algorithm (VFI5), *2*.

[3] Pahareeya, J. (2014). Liver Patient Classification using Intelligence Techniques, *4*(2), 295–299.

[4] Widodo, P. (2014). Rule-Based Classifier to Detect Liver Disease, (1), 71–80.

[5] Parvin, H., Alizadeh, H., & Minaei-bidgoli, B. (2008). MKNN : Modified K-Nearest Neighbor, 22–25.

[6] Parvin, H., Alizadeh, H., & Minati, B. (2010). A Modification on K-Nearest Neighbor Classifier, *10*(14), 37–41.

[7] Zainuddin, S., Hidayat, N., & Andi Subroto, A. (2013). Implementation on Modified K-Nearest Neighbour (Mk-NN) Algorithm on Classifying Soybean Plant Diseases.