Forum Statistika dan Komputasi, Oktober 2010 p : 1-9 Vol 15 No.2
ISSN : 0853-8115

MODELLING INGREDIENT OF JAMU TO PREDICT ITS EFFICACY

Farit Mochamad Afendi[1, 2], Sulistiyani[3], Aki Hirai[1], Md. Altaf-Ul-Amin[1], Hiroki Takahashi[1], Kensuke Nakamura[1] *and* Shigehiko Kanaya[1]
[1] Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
[2] Department of Statistics, Bogor Agricultural University, Bogor, Indonesia
[3] Biopharmaca Research Center, Bogor Agricultural University, Bogor, Indonesia
E-mail : [1] fmafendi@gmail.com

Abstract

Jamu is an Indonesian herbal medicine made from a mixture of several plants. Nowadays, many jamu are produced commercially by many industries in Indonesia. Each producer may have their own jamu formula. However, one is certain; the efficacy of jamu is determined by the composition of the plants used. Thus, it is interesting to model the ingredient of jamu which consist of plants and use it to predict efficacy of jamu. In this analysis, Partial Least Squares Discriminant Analysis (PLSDA) is used in modeling jamu ingredients to predict the efficacy. It is obtained that utilizing $\hat{y}_{ij}$ obtained from PLSDA directly rather than use it to calculate probability of jamu i belong to efficacy j and then use the probability to predict efficacy produces lower False Positive Rate (FPR) in predicting efficacy group.

Keywords: Jamu, PLSDA

INTRODUCTION

Jamu is Indonesian herbal medicine made from a mixture of several plants. Besides useful in curing diseases, jamu also helpful in maintain health capacity ([1]) or even for cosmetic purpose ([2]). In making jamu several plants are selected and mixed so that the efficacy of the concoction obtained is as desired. Traditionally, the plants are chosen by experience from generation to generation and the efficacies of jamu are proven empirically ([2], [3]). In curing the same disease, each ethnic in Indonesia may have their own formulas which depend highly on the plant resources in the region where the ethnic lives ([3], [4]).

Nowadays, many jamu are produced commercially by many industries in Indonesia. Each producer may have their own jamu formula. However, one is certain; the efficacy of jamu is determined by the composition of the plants used ([2]). Thus, it is interesting to model the ingredient of jamu, which consist of plants, and use it to predict the efficacy of jamu.

This paper is organized as follows. In Section 2, the details about the data are given. Basic idea about PLSDA is then discussed in Section 3. This section also explains about the method used in selecting number of components as well as prediction of the efficacy by utilizing prediction of response obtained in PLSDA. Section 4 is prepared for results obtained along with discussion about them. Finally, Section 5 gives the conclusion.

DATA

In this analysis, we focus on commercial jamu in Indonesia, which should be registered and inspected at The National Agency of Drug and Food Control (NA-DFC), so that the safety on people are assured. The information about ingredients of jamu was obtained from this agency, which is provided in their website http://www.pom.go.id/nonpublic/obat_tradisional/default.asp, whereas the information about efficacy of jamu must be obtained from other sources, mainly from the producers. As of February 2010, 6533 jamu produced by local industries in Indonesia were registered at NA-DFC. However, only 3138 jamu could be evaluated for their efficacy. These 3138 jamu were used for our analysis. In total, these 3138 jamu are using 465 plants.

The efficacies of jamu were classified into 9 groups. Then, each jamu was classified into one of these 9 groups. The result is shown in Table 1. Most jamu are useful for gastrointestinal disorders, musculoskeletal and connective tissue disorders, and female reproductive organ problems. All data used in this analysis can be accessed at http://kanaya.naist.jp/jamu/top.jsp.

Table 1. Distribution of jamu according to 9 efficacy groups

Efficacy Group	Frequency of jamu
Disorders of appetite (DOA)	249 (7,9%)
Disorders of mood and behavior (DMB)	22 (0,7%)
Female reproductive organ problems (FML)	398 (12,7%)
Gastrointestinal disorders (GST)	980 (31,2%)
Musculoskeletal and connective tissue disorders (MSC)	840 (26,8%)
Pain/inflammation (PIN)	311 (9,9%)
Respiratory disease (RSP)	107 (3,4%)
Urinary related problems (URI)	72 (2,3%)
Wounds and skin infections (WND)	159 (5,1%)

The details of data structure in this analysis are as follows. The data matrix X in X-block contains status of plant usage in ingredient of jamu. Dimension of matrix X is (N x M), where N is number of jamu and equal to 3138, whereas M is number of plant and equal to 465. Each cell x_{il} is set to 1 if jamu i use plant l, and set to 0 otherwise. On the other hand, the efficacy of jamu acts as response variable Z where Z_i is efficacy of jamu i and it takes 1 out of 9 efficacy groups. However, in PLSDA modeling, this Z variable is then transformed into 9 indicator variables, one for each efficacy group. These 9 indicator variables then perform as Y-block in PLSDA modeling, Thus, dimension of data matrix Y is (N x 9). Each cell y_{ij} is set to 1 if jamu i is classified into efficacy group j, and is set to 0 otherwise. Schematic of data structure is shown in Figure 1.

PLSDA

Partial Least Squares Regression (PLSR) is a regression method which assuming that there are underlying factors in predictors that accounts for most variation in the response. These underlying factors are obtained by maximizing their covariance with the response ([5]). The general underlying model of PLSR ([6]) is

$X = TP^t + E$

$Y = TQ^t + F$

where X is an $n \times m$ matrix of predictors, Y is an $n \times p$ matrix of responses, T is an $n \times c$ matrix of score factors, P and Q are $m \times c$ and $p \times c$ matrix of loading, respectively, and E and F are matrix of error terms.

Although PLSR is not specifically proposed for discrimination among groups. Barker and Rayens ([7]) showed that PLSR can be used for such purpose by connecting PLSR and Linear Discriminant Analysis (LDA) which is called as PLSDA. In PLSDA, the group membership is transformed into a dummy matrix representing group membership. This dummy matrix then performs as response variable in PLSR.

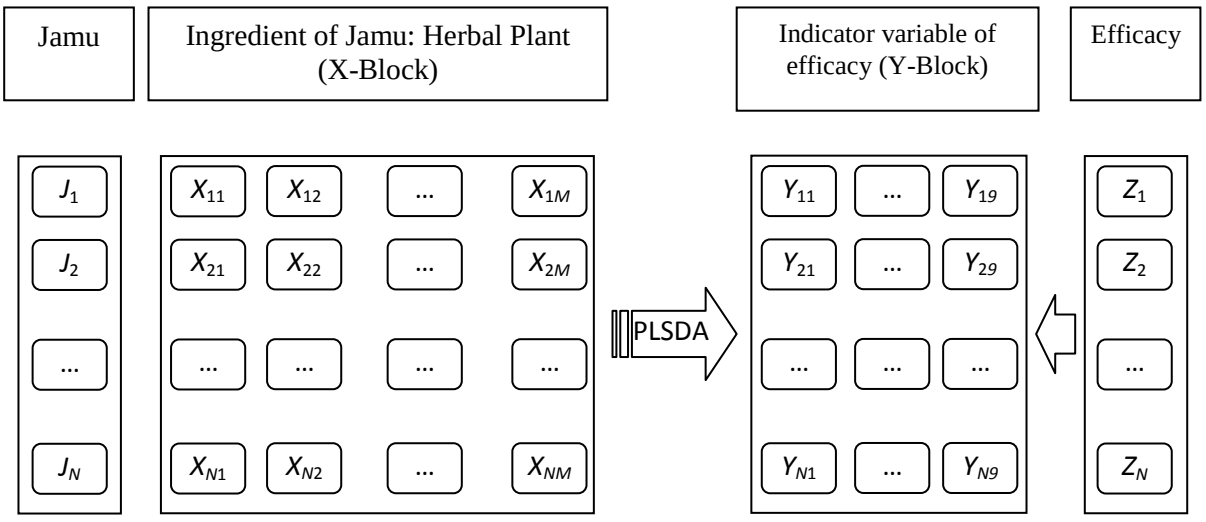


Figure 1 Schematic of data structure used in analysis

Selecting number of components

In this analysis, the number of components in PLSDA is determined by 5-fold cross validation, The steps are as follows,

1. Data are splitted randomly into 5 groups so that each group contains 20% of data. The number of 5 groups or 20% of observations in each group is chosen to ensure that each efficacy group is well represented in each of these 5 groups.
2. One of these 5 groups is chosen as testing data, and the other 4 groups are merged and perform as training data. Then, PLSDA is performed on training data using number of component $k = 1$.
3. The model obtained from step 2 is used talculating this prior probability which are equal o predict *Y*-block value of testing data.
4. Step 2 is repeated by selecting another group as testing data. This step is repeated until all groups are selected as testing data one time.
5. After all groups perform as testing data, Step 2 is again repeated with $k = k+1$. This step is performed until certain number of components.

Let $\hat{y}_{(-i,j)k}$ denotes prediction of response variable *j* using PLSDA model obtained using number of components *k* and without observation *i*. After 5-fold cross validation is performed, Prediction Error Sum of Square (PRESS) using number of components *k* for efficacy group *j* is calculated as

$$PRESS(k)_j = \sum_{i=1}^{n} \left(y_{ij} - \hat{y}_{(-i,j)k}\right)^2,$$

This statistic is then plotted against number of components *k* as of scree plot for eigenvalues.

Prediction of efficacy

In PLSDA, prediction of efficacy can be obtained by utilizing prediction of indicator variable of efficacy $\hat{y}_{ij}$. There are two possibilities in using $\hat{y}_{ij}$ to predict efficacy. The first method is using it directly whereas the second one is utilized it to calculate probability of jamu *i* belong to efficacy *j* and then use the probability to predict efficacy. The prediction of the efficacy for both methods is similar. We assign jamu *i* to efficacy *j* with largest $\hat{y}_{ij}$ for the first method and the largest probability for the second method.

The procedure in utilizing $\hat{y}_{ij}$ to calculate posterior probability of jamu *i* belong to efficacy *j* is as follows. Here Bayes Theorem formula as in ([8]) is used

$$P\left(Class_{ij}\middle|\hat{Y}_{ij},\mu_j,\sigma_j\right) = \frac{P\left(\hat{Y}_{ij}\middle|Class_{ij},\mu_j,\sigma_j\right)P(Class_{ij})}{\sum_{j=1}^{q} P(\hat{Y}_{ij}|Class_{ij},\mu_j,\sigma_j)}$$

In the formula, $P(Class_{ij})$ is prior probability of jamu *i* belong to efficacy *j*. There are two options in calculating this prior probability which are equal across all classes (1/9) and proportional to frequency of each class (see Table 1). Furthermore, $P\left(\hat{Y}_{ij}\middle|Class_{ij},\mu_j,\sigma_j\right)$ is probability of jamu *i* with prediction of indicator variable $\hat{y}_{ij}$ given that jamu *i* belong to efficacy *j* with mean μ_j and standard deviation σ_j. In order to avoid overfitting, ([8]) suggested not to use $\hat{y}_{ij}$ obtained from PLSDA directly but from cross validation procedure as the following,

1. A random sample without replacement is drawn from data as training set to be used for calculation of PLSDA model.
2. The remaining observations are used as testing set. PLSDA model obtained from Step 1 is used to calculate prediction of indicator variable of efficacy for testing set, $\hat{y}_{ij,\text{test}}$.
3. Step 1 and 2 are repeated many times, The predictions of $\hat{y}_{ij,\text{test}}$ are saved across cross validation rounds into $\hat{y}_{ij,\text{cv}}$.

It is assumed ([8]) that $\hat{y}_{ij,\text{cv}}$ is a continuous random variable with the distribution $N(\mu_j, \sigma_j^2)$. The parameters are estimated as

$$\hat{\mu}_j = \frac{1}{t}\sum_{i=1}^{t} \hat{y}_{ij,\text{cv}}$$

$$\hat{\sigma}_j^2 = \frac{1}{t-1}\sum_{i=1}^{t} \left(\hat{y}_{ij,\text{cv}} - \hat{\mu}_j\right)^2,$$

Moreover, class-conditional distributions of $P\left(\hat{Y}_{ij}\middle|Class_{ij},\mu_j,\sigma_j\right)$ used is the probability density function in the form of cumulative distribution function

$$F(x) = P(X \le x).$$

Hence, in posterior probability $P\left(Class_{ij}\middle|\hat{Y}_{ij},\mu_j,\sigma_j\right)$, the likelihood for an observation to belong to class *j* is increasing (to a maximum value of 1) with an increasing $\hat{y}_{ij}$ value.

RESULTS AND DISCUSSION

Selecting number of PLSDA component

PRESS plot of 5-fold cross validation is shown in Figure 2. The plots are relatively constant start from $k = 10$ for all 9 indicator variables. Thus, the number of components is set to 10. Analyzing PLSDA using 10 components we obtain percent variation accounted for predictors and responses as shown in Table 2. Until 10 components, PLSDA can account only 5% variation of predictors. It indicates that there is a weak correlation among usage in jamu of one plant with the other plants. This is reasonable considering there are more than 500 industries producing these 3138 jamu used in this analysis. Although several jamu coming from different producers are useful for the same symptoms, each of their producers has their own jamu formula. This is probably due to, regarding efficacy of jamu, one plant may useful as main

ingredient or as supporting ingredient ([2], [9]). The plants act as supporting ingredient may be replaced with other plants without affecting the efficacy of jamu.

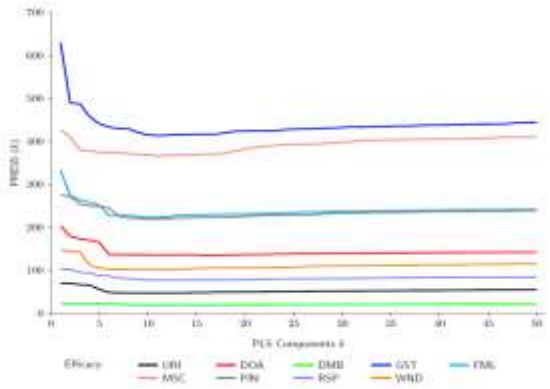


Figure 2 PRESS plot of 5-fold cross validation

Table 2 Percent variation accounted for predictors and responses of PLSDA using k = 10

Number of PLS factors	Percent Variation Accounted For			
	Predictors		Responses	
	Current	Total	Current	Total
1	0,728	0,728	6,874	6,874
2	0,658	1,386	6,871	13,745
3	0,578	1,964	4,474	18,219
4	0,486	2,450	4,672	22,891
5	0,482	2,932	4,486	27,377
6	0,450	3,382	4,555	31,932
7	0,511	3,893	3,103	35,035
8	0,371	4,264	3,239	38,274
9	0,672	4,936	1,215	39,489
10	0,590	5,526	1,015	40,504

Relation of plant with efficacy

Figure 3 shows the distribution of standardized coefficient for each of 9 efficacies. Due to binary nature of y_{ij}, i,e, status of jamu i on efficacy j, then large value of $\hat{y}_{ij}$ will lead to prediction that jamu i is useful for efficacy j. On the other hand, a given plant with positive coefficient will increase the prediction of $\hat{y}_{ij}$; in contrast, a plant with negative coefficient will decrease the prediction of $\hat{y}_{ij}$. Considering these, a plant is assigned as useful for efficacy j if its coefficient on efficacy j is positive. Let B_{lj} be a coefficient of plant l on efficacy j, and U_{lj} be an assignment status of plant l on efficacy j. Thus,

$$U_{lj} = \begin{cases} 1; B_{lj} > 0 \\ 0; \text{otherwise} \end{cases}.$$

Furthermore, if plant l is considered useful for efficacy j then this plant should be used by jamu having efficacy j. To check this, let W_{lj} be plant l usage on efficacy j. W_{lj} is basically the number of jamu with efficacy j and use plant l and calculated as

$$W_{lj} = \sum_{i=1}^{n} X_{il} Y_{ij}.$$

If $U_{lj} = 1$ and $W_{lj} > 0$ then the assignment is called as Hit; on the contrary, if $U_{lj} = 1$ and $W_{lj} = 0$ then the assignment is called as Miss. Table 3 shows summary of Hit-Miss status on this assignment.

Table 3 informs that there are many plants categorized as Miss. It means that these plants are assigned as useful for certain efficacy, because they have positive coefficient on that efficacy, but in fact there are no jamu with that efficacy uses the plants. However, by exploring range of coefficient values as shown in Figure 4, coefficient values of plants categorized as Miss are insignificant from 0.

Table 3 Assignment status of plant to efficacy using positive value of coefficient

Efficacy	$U_{lj} = 1$		$U_{lj} = 0$
	Hit	Miss	
URI	48	32	385
DOA	94	8	363
DMB	35	72	358
GST	149	8	308
FML	115	50	300
MSC	172	6	287
PIN	113	25	327
RSP	62	81	322
WND	86	11	368

In order to reduce the number of Miss, further improvement of assigning plant to efficacy is conducted as follow. Note that each plant has 9 coefficients, one for each efficacy. Rather than assigning plant on each efficacy with positive coefficient, the new assignment will assign the plant only on efficacy with largest coefficient. Thus, if V_{lj} denotes the new assignment status of plant l to efficacy j, then

$$V_{lj} = \begin{cases} 1; B_{lj} = \max_{j}(B_{lj}) \\ 0; \text{otherwise} \end{cases},$$

For this new assignment, if $V_{lj} = 1$ and $W_{lj} > 0$ then the assignment is called as Hit; whereas if $V_{lj} = 1$ and $W_{lj} = 0$ then the assignment is called as Miss. Table 4 shows summary of Hit-Miss status for this new assignment. It is obtained that the number of Miss for the new assignment is 5 plants out of 465 plants.

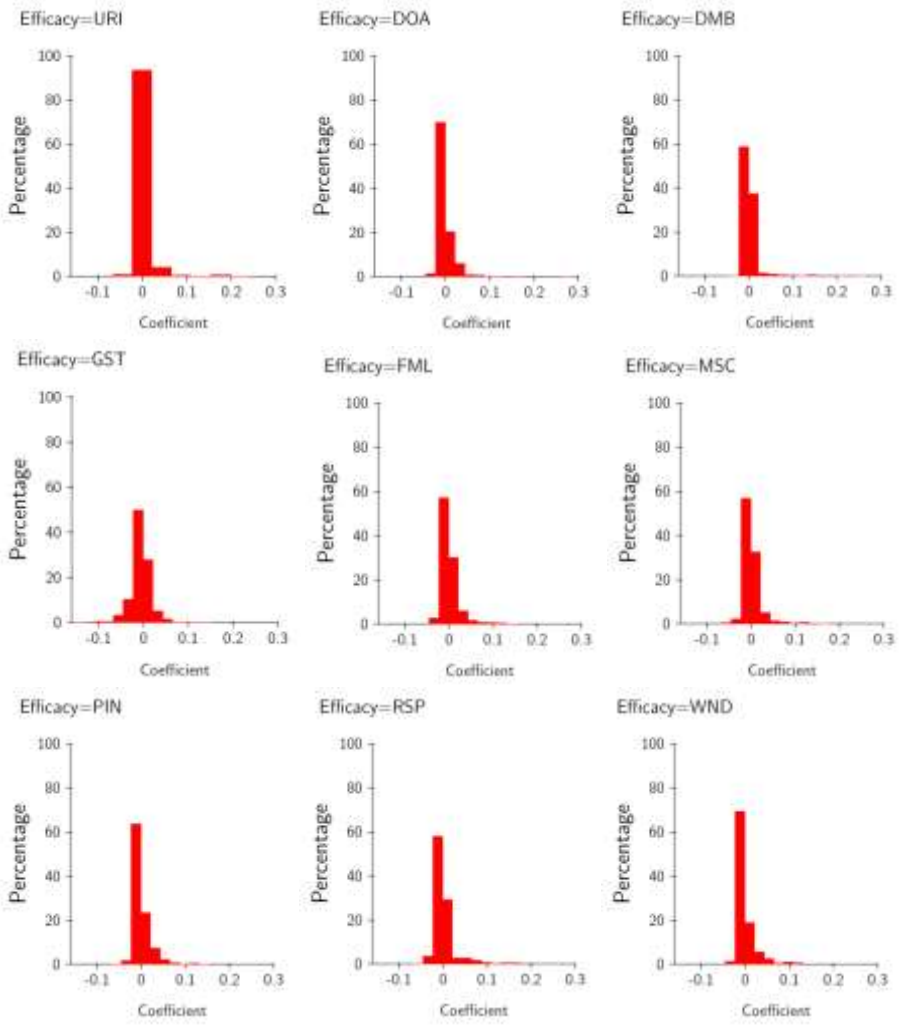


Figure 3 Distribution of standardized coefficients for each response

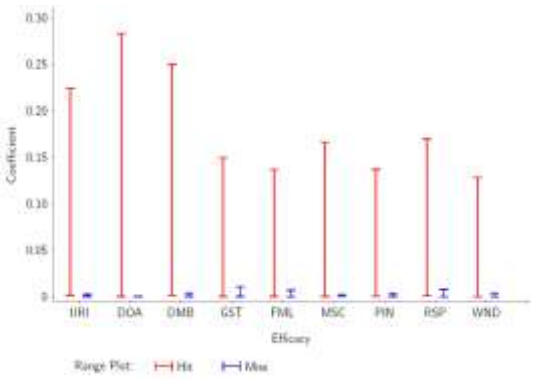


Figure 4 Range plot of standardized coefficient values for each of efficacies

Prediction of efficacy

The matrix T or scores of predictors in PLS can be regarded as summary of predictors which contain useful information in predicting responses. Plot among these scores can be used to explore PLS performance in predicting responses. The plots among the first three predictors' scores are shown in Figure 5. It is obvious that many points from different efficacies are overlapping. This overlapping between points from different efficacy is also obtained over other scores (results not shown). This is because many plants are used for more than one efficacy. Then each scores obtained are not specific for certain efficacy. Hence all 10 scores must be used simultaneously in predicting efficacy.

Table 4 Hit-Miss status of assignment of plants to efficacy using maximum coefficient

Efficacy	Hit	Miss
URI	23	0
DOA	45	0
DMB	13	0
GST	82	1
FML	61	4
MSC	94	0
PIN	69	0
RSP	31	0
WND	42	0

From PLSDA model we obtain 9 $\hat{y}_{ij}$, one for each indicator variable of efficacy. Distribution of $\hat{y}_{ij}$ against indicator variable for each efficacy is

shown in Figure 6, From Figure 6 it is obtained that the center of $\hat{y}_{ij}$ between $Y_j = 0$ and $Y_j = 1$ are well separated for all efficacies. Employing T-Test for each efficacy in testing the equality of the center of $\hat{y}_{ij}$ between $Y_j = 0$ and $Y_j = 1$ also support that both center are well separated (Table 5).

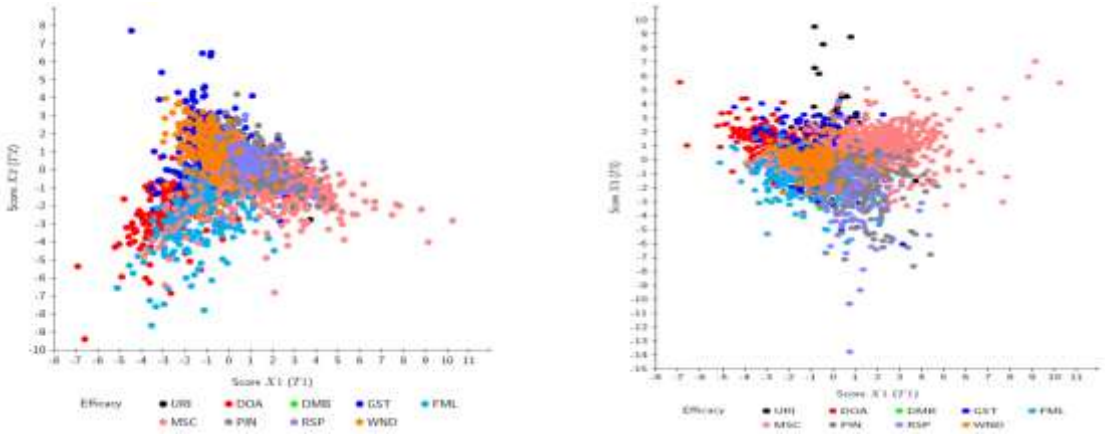
Although the center of $\hat{y}_{ij}$ between $Y_j = 0$ and $Y_j = 1$ are well separated, if we examine each efficacy in Figure 6 it is obvious that there are overlapping region where both $Y_j = 0$ and $Y_j = 1$ have the same $\hat{y}_{ij}$. This finding is similar with predictor's scores in Figure 5. However, Area Under Curve (AUC) statistic of ROC Curve of all efficacies (Table 5 **Error! Reference source not found.**) indicate that the prediction of indicator variable of efficacy $\hat{y}_{ij}$ are a good candidate in discriminating $Y_j = 0$ and $Y_j = 1$.

Distribution of $\hat{y}_{ij,\text{cv}}$ using 200 rounds of cross validation along with its normal curve is shown in Figure 7. It is obtained that distributions of $\hat{y}_{ij,\text{cv}}$ for all efficacies are not normal. Hence, in this analysis we make two options regarding distribution of $\hat{y}_{ij.\text{cv}}$. The first option still assuming that the distribution of $\hat{y}_{ij,\text{cv}}$ is normal whereas the second option will use empirical distribution obtained from cross validation as distribution of $\hat{y}_{ij.\text{cv}}$.

Table 5 T test results in comparing mean of $\hat{y}_{ij}$ between $Y_j = 1$ and $Y_j = 0$ and AUC for ROC curve

Efficacy	T-Test		AUC
	T value	P value	
URI	-10,63	<,0001	0,978
DOA	-24,27	<,0001	0,947
DMB	-3,77	0,0011	0,983
GST	-55,34	<,0001	0,933
FML	-29,08	<,0001	0,932
MSC	-40,92	<,0001	0,913
PIN	-19,06	<,0001	0,916
RSP	-13,27	<,0001	0,954
WND	-14,48	<,0001	0,950

By using maximum $\hat{y}_{ij}$ and probability method, prediction of efficacy is conducted with the result of FPR is shown in Table 6. It is obtained that, in probability method, prediction using equal prior is better than proportional prior. This is due to the frequency is distorted to two efficacies, namely GST and MSC, which make prediction using proportional prior also distorted to these two efficacies (Table 7). On the other hand, still in probability method, assuming normal distribution for class-conditional distribution lead to a better prediction than using empirical distribution. It is because there are outliers in the distribution of $\hat{y}_{ij,\text{cv}}$ (Figure 7).



(a) (b)

Figure 5 Predictors' scores plot for: (a) component 1 vs component 2, and (b) component 1 vs component 3

Furthermore, combination of equal prior and normal distribution in class conditional distribution, although better than any other combination in probability method, produces larger FPR compare to maximum $\hat{y}_{ij}$ method. This is because equal prior is not informative as a prior whereas assumption of normal distribution for class conditional distribution is violated (Figure 7). Hence, maximum $\hat{y}_{ij}$ method then is used for prediction of the efficacy with the result of confusion matrix is shown in Table 8. The correct classifications for each efficacy are range from 22,7% for efficacy DMB until 89,8% for efficacy GST. Error in prediction of the efficacy occurs because plants usage in jamu is not unique for certain efficacy. Many plants are used for more than one efficacy. Therefore, in future works, pharmaceutical activities of plants will be used in improving the model. It is expected that by adding this new information the function of plants in jamu are clearer, although their usage are not unique for certain efficacy, and the prediction of the efficacy of jamu will be better.

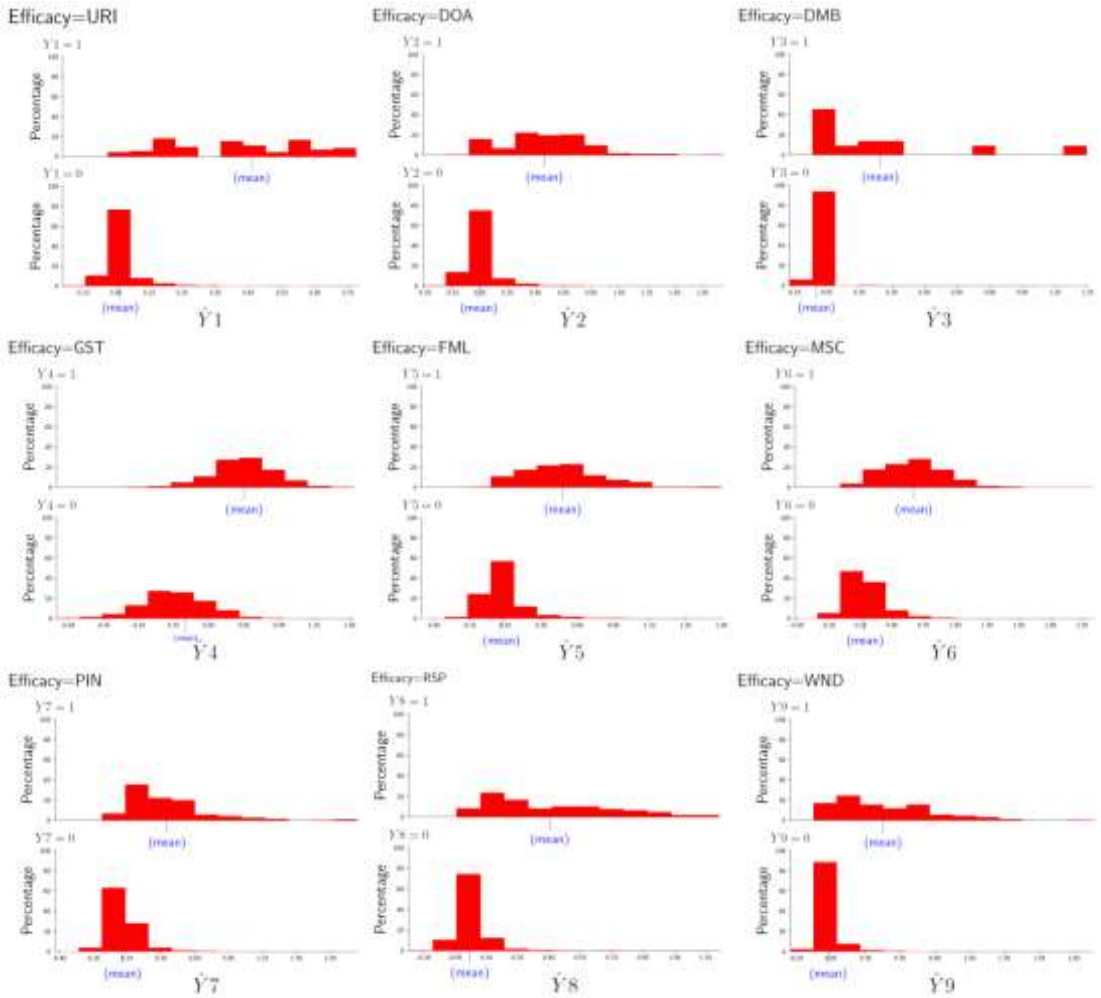


Figure 6 Distribution of $\hat{y}_{ij}$ against indicator variable of efficacy for all 9 efficacies

Table 6 FPR of the prediction of the efficacy using maximum $\hat{y}_{ij}$ and using probability method

Prediction method	FPR
Maximum $\hat{y}_{ij}$	0,284
Probability	
• Equal prior – Normal distribution (Eq—N)	0,328
• Equal prior – Empirical distribution (Eq—E)	0,439
• Proportional prior – Normal distribution (Pr—N)	0,417
• Proportional prior – Empirical distribution (Pr—E)	0,434

Table 7 Result of prediction of efficacy using maximum $\hat{y}_{ij}$ method and probability method, Here TC and CC means Total Classification and Correct Classification, respectively.

Efficacy	Observed	Maximum $\hat{y}_{ij}$ method		Probability method							
				Eq—N		Eq—E		Pr—N		Pr—E	
		TC	CC	TC	CC	TC	CC	TC	CC	TC	CC
URI	72	54	39	184	67	406	69	0	0	0	0
DOA	249	204	164	295	188	277	182	20	15	13	8
DMB	22	6	5	85	19	318	22	0	0	0	0
GST	980	1296	880	790	672	518	484	1634	919	1580	906
FML	398	376	266	415	274	375	256	285	175	256	136
MSC	840	876	638	617	514	426	373	1128	674	1238	695
PIN	311	171	133	282	180	265	171	69	44	50	30
RSP	107	65	52	285	91	327	93	0	0	0	0
WND	159	90	71	185	105	226	111	2	2	1	1

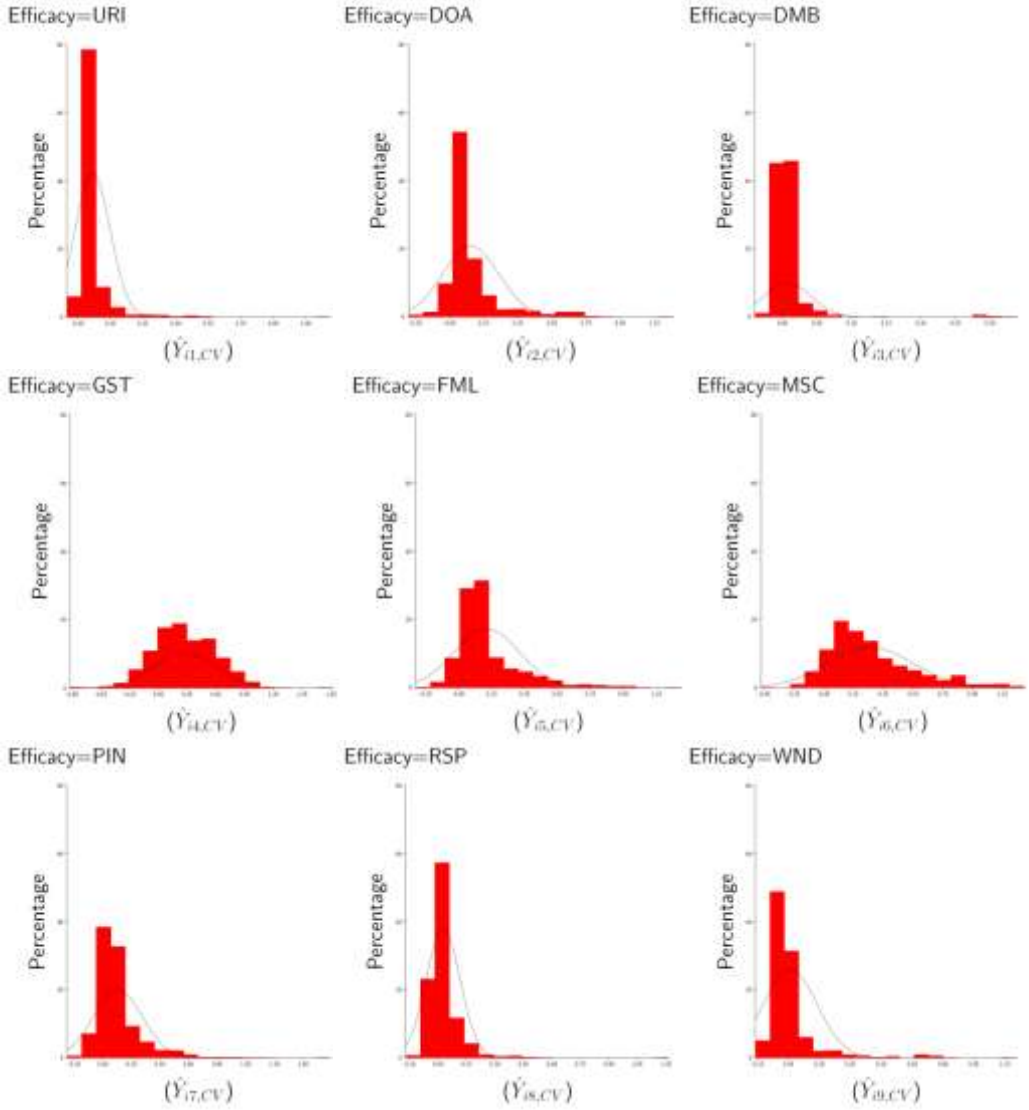


Figure 7 Distribution of $\hat{\boldsymbol{y}}_{\boldsymbol{ij},\boldsymbol{cv}}$ along with normal curve for all efficacy

Table 8 Confusion matrix of prediction of efficacy using maximum $\hat{\boldsymbol{y}}_{\boldsymbol{ij}}$ method

Observed	Predicted									Total
	URI	DOA	DMB	GST	FML	MSC	PIN	RSP	WND	
URI	39	0	0	21	2	10	0	0	0	72
DOA	0	164	0	29	36	18	0	0	2	249
DMB	0	1	5	10	0	3	1	2	0	22
GST	3	17	0	880	12	46	9	6	7	980
FML	0	13	0	61	266	50	5	1	2	398
MSC	6	6	1	127	41	638	16	0	5	840
PIN	1	0	0	90	4	77	133	4	2	311
RSP	3	0	0	21	4	23	3	52	1	107
WND	2	3	0	57	11	11	4	0	71	159
Total	54	204	6	1296	376	876	171	65	90	3138

CONCLUSION

In this analysis, PLSDA is used in modeling jamu ingredients to predict the efficacy. It is obtained that utilizing $\hat{\boldsymbol{y}}_{\boldsymbol{ij}}$ obtained from PLSDA directly rather than use it to calculate probability of jamu *i* belong to efficacy *j* and then use the probability to predict efficacy produces lower FPR in predicting efficacy group.

REFERENCES

1] S. J. Beers, ”Jamu: The Ancient Indonesian Art of Healing”, Vermont: Periplus Editions (HK) Ltd, 2001.

[2] S. Pramono, ”Jamu in Indonesian Daily Life and Industry”, Toyama: Institute of Natural Medicine, University of Toyama, 2007.

[3] I. K. Adnyana, A. A. Soemardji, ”Evaluation of Pharmacological Efficacy of Jamu Medicine”, Toyama: Institute of Natural Medicine, University of Toyama, 2007.

[4] H. M. Sangat, E. A. M. Zuhud, E. K. Damayanti, ”Kamus Penyakit dan Tumbuhan Obat Indonesia [Etnofitomedika]”, Jakarta: Yayasan Obor Indonesia, 2000.

[5] S. Wold, M. Sjöström, L. Eriksson, ”PLS-regression: a basic tool of chemometrics”, *Chemometrics and Intelligent Laboratory Systems*, Vol. 58, pp. 109-130, 2001.

[6] A. Boulesteix, K. Strimmer, ”Partial least squares: a versatile tool for the analysis of high-dimensional genomic data”, *Briefings in Bioinformatics*, Vol. 8, pp. 32-44, 2006.

[7] M. Barker, W. Rayens, ”Partial least squares for discrimination”, *J. Chemometrics*, Vol. 17, pp. 166-173, 2003.

[8] M. Bylesjö, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, J. Trygg, ”OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification”, *J. Chemometrics*, Vol. 20, pp. 341-351, 2006.

[9] Anonim, “Herbal Indonesia Berkhasiat: Bukti Ilmiah dan Cara Racik”, Prepared by Redaksi Trubus, Depok: PT Trubus Swadaya, 2009.