ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 650-654
doi: 10.18178/wcse.2016.06.114

Support Vector Machine OVA-RFE Approach for Finding the Significant Plants of Jamu

Aries Fitriawan [1, 2], Ito Wasito [1], Wisnu Ananta Kusuma [2, 4 +], and Rudi Heryanto [3, 4]

[1] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

[2] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Bogor, Indonesia

[3] Department of Chemistry, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Bogor, Indonesia

[4] Biopharmaca Research Center, Bogor Agricultural University, Bogor, Indonesia

Abstract. Jamu medicines are popular traditional medicines from Indonesia. Jamu made from a mixture of several plants. Jamu formula are based on empirical data and personal experiences, so it needs to systemize the formulation of Jamu and develop basic scientific principles of Jamu for Indonesian Healthcare System. The purpose of this research is to find the Jamu plants that have the most significant effect on the diseases. We proposed a feature selection approach using SVM OVA-RFE. We also added the previous Jamu feature selection research using K-Means and PLS-DA for comparison. The SVM OVA-RFE method successfully reduced the data dimension into 3085 of Jamu samples and 238 species of plants. The result from SVM classification using OVA-RFE outperform the previous researches.

Keywords: Jamu, feature selection, machine learning, recursive feature elimination, SVM OVA-RFE.

1. Introduction

The number of medicinal plants is estimated to be 40,000 to 70,000 around the world [1]. In some countries, these plants are used as a mixtured herbal medicine, for example Traditional Chinese Medicine (China), Kampo (Japan), Ayurveda (Indian), and Jamu (Indonesia). The use of Jamu as alternative medicine is increasing, since its safety and efficacy given by Jamu [2]. Combinations of extracts from various plants had a better synergic interactions than those that used the singularly extract [3]. Therefore, Jamu have many variations of formula in every region in Indonesia. These Jamu formula are based on natural ingredients that exist in those regions [4]. However, although Jamu has been empirically proven to cure some diseases, there is no sufficient scientific evidence to explain the relationship between the composition of plants and their efficacy. In accordance with the policy from Ministry of Health, Republic of Indonesia in 2010 about scientification of Jamu, it is required to systemize the formulations and develop basic scientific principles of Jamu to support the Indonesian Healthcare System [5].

The systematic effort to find the relationship between plants in Jamu were conducted in many studies. Afendi *et al.* [6] used Partial Least Square-Discriminant Analysis (PLS-DA) to develop a model to classify efficacy of Jamu formulas. Results showed that the accuracy from 10-fold cross validation evaluation was 71.6% [6], However, accuracy for the selection feature didn’t make any significant result (70.64%) [7]. Machine learning approaches are also done by Fitriawan *et al.* [8] and Puspita *et al.* [9]. Fitriawan *et al.* used Support Vector Machine (SVM) to seek Jamu classification model based on data from previous researches [6] [7]. The research [8] resulted a better accuracy for the data that has been through a filtering process (95.7%),

[+] Corresponding author. Tel.: +6285313899144
E-mail address: ananta@ipb.ac.id.

but this research have not been able to find plants that have a significant effect to a disease. Meanwhile, Puspita *et al*. [9] used K-Means feature selection approach to find the best set of significant Jamu plants.

This topic still has a challenging task which is searching for significant Jamu plants. In the field of machine learning, searching for significant features problem can be solved by feature elimination method. In this research, we use SVM-based feature elimination. For multiclass Jamu classification, we use one-versus-all SVM-RFE (SVM OVA-RFE) [10]. Results of this method will be compared with the result from previous researches. In order to get a fair comparison, we used the same dataset as the one used in PLS-DA [7]. The output from this research is expected to become a baseline for finding the Jamu's core plants that have a direct effect for the diseases in the future work.

2. Methodology

This research was developed under the methodology (see Fig. 1). This methodology aimed to develop and evaluate the selection feature method for finding core plants of Jamu.

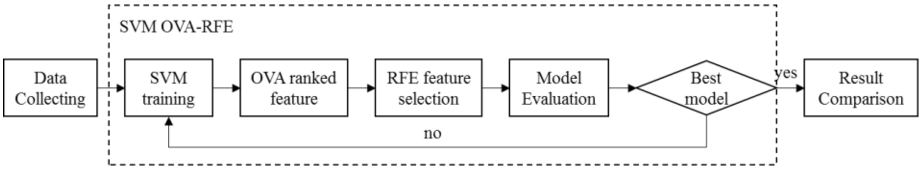
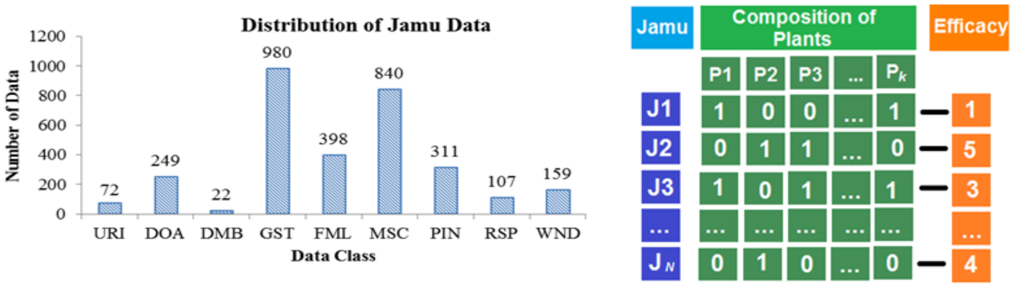


Fig. 1: The scheme for the research method used in developing feature selection for Jamu's plants and comparing the result from SVM OVA-RFE.

3. Data

We used the Jamu dataset from Afendi *et al.* [6]. The data consisted of 3138 Jamu samples that listed in the National Agency of Drug and Food Control (NADFC), Indonesia. Most of the samples contained one to sixteen plants which were taken from 465 plants.

The data contained nine classes that represent the general efficacy of Jamu medicine against the diseases. These class division are based on the division of the *International Classification of Diseases ver. 10* after through a simplification process in previous research [6]. Those are urinary related problems (URI), disorder of apetite (DOA), disorder of mood and behavior (DMB), gastrointestinal disorders (GST), female reproductive organ problems (FML), muskuloskeletal and connective tissue disorders (MSC), pain and inflammation (PIN), respiratory disease (RSP), and wounds and skin infections (WND). The number of samples of each dataset assigned to each effects can be seen in Fig. 2 (A). We could generate feature vectors from the dataset. Features were represented by composition of plants in Jamu samples (See Fig. 2 (B)). Composition of plants was defined by assigning binary value to each plants. If a certain plant is included into a Jamu sample then the value of this plant is assigned to 1, otherwise the value is 0.



(A) (B)

Fig. 2: (A) Data distribution used in this research. (B) Representation of data relationships between Jamu, plants and Jamu efficacy [8].

4. Support Vector Machine–Recursive Feature Elimination for Multiclass Problem

In SVM-RFE algorithm, each iteration will include steps as follows [11]:

Step 1. Training the SVM classifier.

Step 2. Calculating weight vector $w = \sum_k a_k \, y_k \, x_k$, where x_k is the data vector of a sample k, y_k[-1, +1] encodes the class lable of sample k, and a_k is calculated from the training set.

Step 3. Calculate the ranking score $c = (w)^2$

Step 4. Seeking the feature with smallest ranking score $f = argmin(c)$.

Step 5. Eliminating the feature with $c = f$

SVM-RFE is originally proposed for binary problems. So in the case of multiclass classification, another approach is required, one of which is One-versus-all (OVA) method. Zhou and Tack [10] employed the extension of SVM-RFE called OVA-RFE. For k-class problem (k≥3), k binary SVM classifiers are constructed using OVA method. For the k-th binary classification problem, SVM-RFE is carried out to identify a feature subset S_k for class k against all other classes. After k feature subsets are selected, the final selected subset for the whole multiclass problem is the combination of all the k subsets.

5. Evalution Method

Evaluation is done by calculating the accuracy performance of SVM OVA-RFE method. This research will calculate accuracy from 10-fold cross validation training.

$$accuracy = \frac{true\ positives + true\ negatives}{number\ of\ data}$$

6. Experiments Result and Disscussion

This research perform SVM OVA-RFE which is done by eliminating every 10 of features. This experiment aims to find how many significant feature in these data and get the best accuracy between the range of data. This experiment compare the result from 4 different Kernel Function (linear, 2 degree polynomial, 3 degree polynomial, and RBF) by using 10-fold cross validation training. The comparable chart can be seen at Fig. 3.

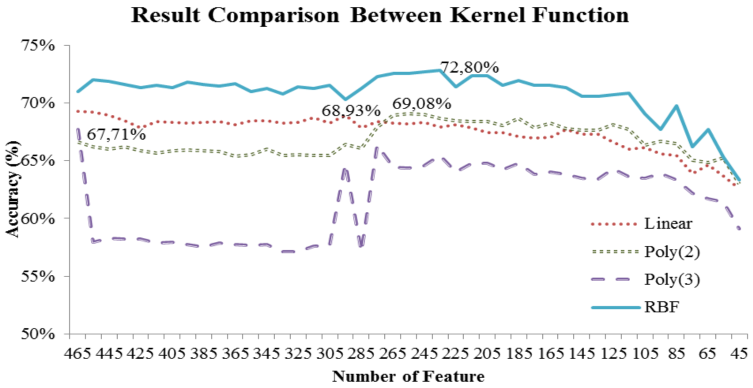
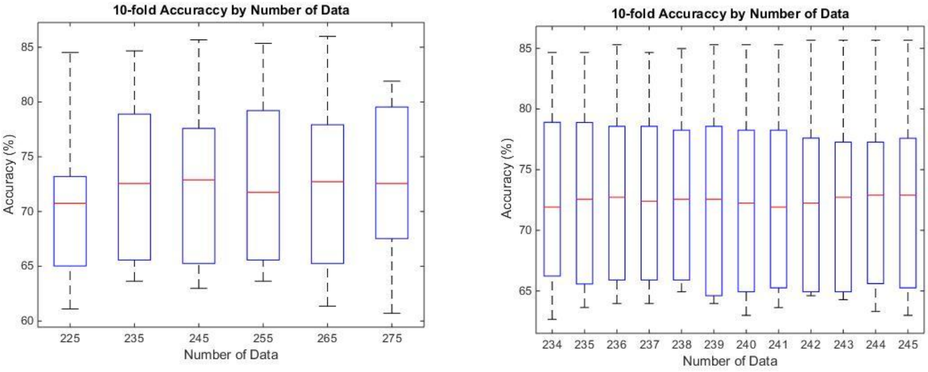


Fig. 3: Comparison chart between kernel function.

As we can see at Fig. 3, Radial Basis Function (RBF) kernel outperform the other Kernel function in accuracy. The result shows that RBF kernel have the highest accuracy 72,80% with 235 features; 2-degree Polynomial kernel accuracy 69,08% with 255 features; Linear kernel accuracy 68,93% with 295 features; and 3-degree Polynomial kernel accuracy 67,71% with 465 features. From the result, we conclude RBF Kernel has better performance than the other Kernel Function.

To avoiding the local maxima problem from this method, the boxplot analysis are presented to see the distribution of accuracy in each fold of cross validation (Fig. 4). Fig. 4(A) contain the result from 225 until 275 features with 10 major unit, while Fig. 4(B) have more detail of distribution between 234 and 245 features. The best accuracy from 10-fold cross validation average is obtained 73,06% ± 6.80% from 238 best features. This best result is obtained after parameter tuning. The parameter tuning is done by using grid search method. This grid search method are done by using grid.py tools from LibSVM library [12]. The detail of output from grid search can be seen at Fig. 5. From Fig. 5, we get accuracy 73.06% using SVM

RBF Kernel with parameter $c = 2$ and $\gamma = 0.25$. We also reduce the number of Jamu sample to 3085 because there is a sample that didn't have any plants in its formula.



(A) (B)

Fig. 4: (A) Boxplot for accuracy distribution between 225 and 275 features. (B) Boxplot for accuracy distribution between 234 and 245 features.

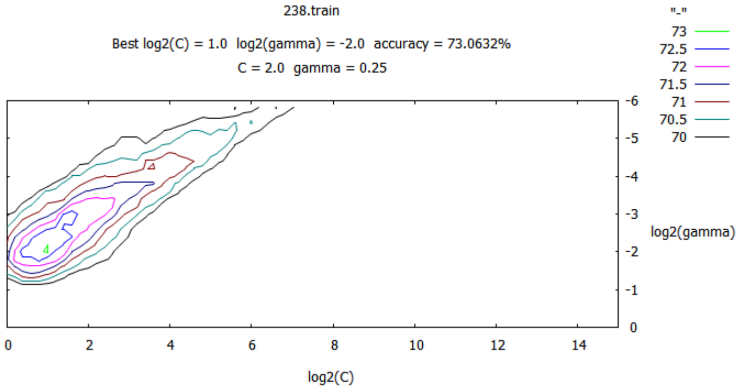


Fig. 5: Grid search result chart for RBF Kernel SVM from best feature selection data.

We also compare the result with the previous research [7][9]. For a fair comparison, this research will put the first data treatment from Afendi *et al.* [7] into the comparison table with the other method (PLS-DA [7] and K-Means [9]). The comparison between the method can be seen at Table 1. From Table 1, SVM OVA-RFE outperform the other methods.

Table 1. Comparison Between Methods

Method	Features	Accuracy (%)
PLS-DA [7]	231	70.64
SVM + K-Means [9]	236	71.50
SVM OVA-RFE	238	73.06

7. Conclusion

This research has successfully applied the recursive feature selection for Jamu plants. We also compare the accuracy of Jamu formulation based on efficacy classification between SVM K-Means feature selection, PLS-DA, and our method, SVM OVA-RFE. The accuracy of our method is comparable to those of PLS-DA and SVM K-Means feature selection. Our method obtained 72,06% accuracy for SVM OVA-RFE. This accuracy is slightly higher than PLS-DA or SVM K-Means.

8. Acknowledgements

The authors wish to thank Dr. Farit Mochamad Afendi for his help in collecting the data of Indonesian Jamu from National Agency of Drug and Food Control (NADFC) Republic of Indonesia. The authors also wish to thank Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan, LPDP), Ministry of Finance, Republic of Indonesia that has funded this publication.

9. References

[1] J. Li, Y. Jiang, R. Fan. Recognition of Biological Signal Mixed Based on Wavelet Analysis. In: Y. Jiang, et al (eds.). *Proc. of UK-China Sports Engineering Workshop*. Liverpool: World Academic Union. 2007, pp. 1-8. (Use “References” Style)

[2] R. Verporte, H. K. Kim, and Y. H. Choi, Plants as source of medicines. In: R. J. Boger, L. E. Craker, and D. Lange (Eds.). *Medicinal and Aromatic Plants*. 2006, Chapter 19, pp. 261–273.

[3] G. B. Mahady Global harmonization of herbal health claim. *J. of Nutr*. 2001, 131: 1120S-1123S.

[4] L. S. Adam, P. S. Navindra, L. H. Mary, C. Catherine, H. David. Analysis of the interactions of botanical extract combinations againts the viability of prostate cancer cell lines. *Evid Based Complement Alternat Med.* 2006, 3(1): 117-124.

[5] S. H, Wijaya, H. Husnawati, F. M. Afendi, et al., Supervised Clustering Based on DPClusO: Prediction of Plant-Disease Relations Using Jamu Formulas of KNApSAcK Database. *BioMed Research International*. 2014, vol. 2014, Article ID 831751, 15 pages, 2014. doi:10.1155/2014/831751.

[6] F. M. Afendi, L. K. Darusman, A. Hirai, M. A. Amin, H. Takahashi, K. Nakamura, S. Kanaya, “System biology approach for elucidating the relationship between Indonesia herbal plants and the efficacy of Jamu (Published Conference Proceedings style),” in *IEEE International Conference on Data Mining Workshops*, Sydney, Australia, Des 2010, Sydney (AU) : Conference Publishing Services, pp. 661-668.

[7] F. M. Afendi, L. K. Darusman, M. Fukuyama, M. Altaf-Ul-Amin and S. Kanaya. “A bootstrapping approach for investigating the consistency of assignment of plants to Jamu efficacy by PLS-DA model,” *Malaysian Journal of Mathematical Sciences.* 2012, vol. 6, no.2, pp. 147-164.

[8] A. Fitriawan, W. A. Kusuma, R. Heryanto. "A classification system for Jamu efficacy based on formula using Support Vector Machine*" 2013 International Conference on Advanced Computer Science and Information Systems* (ICACSIS). 28-29 Sept. 2013. pp.291-295. doi: 10.1109/ICACSIS.2013.6761591

[9] M. N. Puspita, W. A. Kusuma, A. Kustiyo, R. Heryanto. “A classification system for Jamu efficacy based on formula using Support Vector Machine and K-Means Algorithm as Feature Selection*" 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Depok, 2015, pp. 215-220. doi: 10.1109/ICACSIS.2015.7415176

[10] X. Zhou, D. P. Tuck. "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data*" Bioinformatics*. 2007, vol 23 no.9, pp. 1106-1114.

[11] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines", *Machine Learning*. 2002, vol. 46, nos. 1-3, pp. 389-422.

[12] C.W. Hsu, C.C. Chang and C. J. Lin. (2012, Des 1). *A practical guide to support vector classification*. Technical Report, National Taiwan University [Online]. Available : https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/s vmguide.pdf