Implementation of hybrid sampling technique for predicting active compound and protein interaction in unbalanced dataset

W A Kusuma¹,³, A S Rahmi¹, and R Heryanto²,³

¹Department of Computer Science, Bogor Agricultural University, Bogor, West Java, Indonesia e-mail: ananta@apps.ipb.ac.id
²Department of Chemistry, Bogor Agricultural University, Bogor, West Java, Indonesia
³Tropical Biopharma Research Center, Bogor Agricultural University, Bogor, West Java, Indonesia

Abstract. Indonesia Jamu Herbs (Ijah) web server aims to predict Jamu efficacy based on interaction between active compound and disease’s protein. However, the interaction between compound and protein data is unbalance since there are many unknown interactions between active compounds and protein target. Thus, the prediction result is still not optimal. In this research, the hybrid sampling technique, combining complementary fuzzy support vector machine (CMTFSVM) and synthetic minority oversampling technique (SMOTE) was used to handle imbalanced data interaction between active compound and protein for Ijah, web server to predict candidate Jamu formula for certain disease. Performance was measured using geometric mean (Gmean), area under curve (AUC), and accuracy. The evaluation results showed that the hybrid sampling technique could increase the instance of minority class three times. Moreover, the prediction model could obtain the value of 0.8346, 0.6812, and 0.5319 for accuracy, Gmean, and AUC, respectively.

Keywords: Indonesia Jamu Herbs (Ijah), efficacy, active compound, fuzzy support vector machine, SMOTE

1. Introduction
Jamu is Indonesian herbal medicine which is used to maintain health and to treat a certain disease[1]. The mixture of some plants composing Jamu can treat certain health problems. The knowledge about Jamu is derived based on local society experience and culture[2]. Study about the association between plant and disease conducted to find jamu formula using the relation between active compound in plant and protein in disease[3].

Implementation of bipartite local model network–based interaction–profile inferring (BLMNII) can predict the relation between active compound and protein to create jamu formula which is implemented to web application called Indonesia Jamu Herbs (Ijah)[4]. BLMNII uses Support Vector Machine (SVM) classifier and the information of active compound and protein interaction to predict candidate jamu formula for a certain disease. If there are no information of protein and active compound interaction, calculation of the class label uses the most similar active compound and protein from itself. Prediction without interaction information was the worst cases in BLMNII with O(n²). The implementation in [4] used unbalanced data where the information between active compound and
protein interaction was only 0.0001%. Prediction with unbalanced data in BLMNII used similarity of other data to predict Jamu candidate, this situation was considered as the worst case of BLMNII. BLMNII can not handle unbalanced data therefore there should be a method to balancing the active compound and protein interaction data in order to minimize the bias of classifier decision.

There are sampling techniques used to handle many unbalanced data cases such as undersampling, oversampling, and hybrid sampling[5,6]. Undersampling technique is conducted by decreasing the number of majority class data while oversampling technique aims to increase the number of data in minority class to achieve more balanced data. The hybrid sampling is the combination of undersampling and oversampling techniques. The Complementary Fuzzy Support Vector Machine (CMTFSVM) combined with Synthetic Minority Oversampling Technique (SMOTE) is an example of hybrid sampling technique which could handle unbalanced data effectively[6]. CMTFSVM is an undersampling technique which applies complementer (CMT) concept with truth model and falsity model. CMTFSVM uses fuzzy membership values to representing the importance of data to their class. SMOTE as oversampling technique, was combined to increase the number of minority class and to create more balanced dataset.

This study aims to implement hybrid sampling technique using CMTFSVM dan SMOTE to handle unbalanced data interaction between active compound and protein in Ijah. Implementation method referred to the research conducted by Pruengkarn[6].

2. Materials and methods

2.1. Dataset

Dataset used in this research referred to the research data used by Pruengkarn[6], Kurnia[4], and by webring from PubChem. Pruengkarn’s dataset was consisted of three dataset including German, Yeast3, and Glass5 dataset. All dataset was binary dataset which only had two classes: positive class (minority class) and negative class (majority class). The German dataset is used to identify credit card fraud, where the positive class represents fraud occurrence. Moreover, the Yeast3 dataset cluster yeast into ME3 (positive class) and others (negative class). Lastly, the Glass5 dataset tried to identify glass source based on the chemical information. There were two classes in the Glass5 dataset, including container glass (positive class) and non container glass (negative class). Table 1 summarises characteristics of these three datasets. The summarization was conducted according to attributes such as the number of total instances (#instance), the majority class percentage (%major), and the minority percentage (%minor).

| Dataset  | #instance | %major | %minor |
|----------|-----------|--------|--------|
| German   | 1 000     | 70     | 30     |
| Yeast3   | 1 484     | 89.02  | 10.98  |
| Glass5   | 214       | 95.79  | 4.21   |

Kurnia’[4] dataset consists of list of compound, list of protein, list of compound-protein interaction, and amino-acid sequence taken from Uniprot. Every compound has chemicals characteristics called compound descriptor. In this study, compound descriptor was referred to PubChem. Compound descriptors were crawled from PubChem using Compound ID (CID) taken from list of compound. Moreover, protein descriptors were obtained based on feature calculations. For instance, Amino acid composition (AAC) is one type of protein descriptors, it describe a protein as the proportion of each amino acid divided by total amino acid that construct the protein. Protein feature server (Profeat)[7] provides protein descriptor using amino acid composition (AAC) feature.

Protein descriptor and compound descriptor are the variable of list of compound-protein interaction used in Ijah dataset to predict jamu formula. The value of compound-protein interaction becomes the label of class, where the positive class represents the existence of interaction between compound and
protein, the negative class, otherwise. The dataset generated by pairing of protein descriptor and compound descriptor, that labelled based on list of compound-protein interaction.

2.2. Methods
Figure 1 illustrates the methods. The experiment was divided into two part. Part (a) used the three dataset from Pruengkarn[6] with FSVM as classifier. The Geometric Mean (Gmean), Area Under Curve (AUC), and Accuracy was used to evaluate the results. The purpose of stage (a) is to validate the implementation of hybrid sampling technique which refer to Pruengkarn’s study[6].

![Figure 1. Experimental process.](image)

Stage (b) is the implementation of hybrid sampling technique using Ijah dataset which used compound descriptor and protein descriptor. Ijah dataset is a matrix which pairing each protein descriptor with each compound descriptor and labelled by compound protein interaction (CPI) data. The evaluation was conducted using SVM as classifier to get geometric mean (Gmean) and the accuracy of model. The balanced dataset was taken from hybrid sampling technique filtered by CPI similarity and was used as BLMNII input.

The implementation of Hybrid sampling technique can be seen in figure 2. The hybrid sampling technique was implemented using CMTFSVM combining with SMOTE. The order of implementation was referred to study performed by Pruengkarn[6]. The evaluation was conducted by comparing Gmean and AUC of using undersampling, oversampling, and hybrid sampling technique in different order using Support Vector Machine (SVM), Fuzzy Support Vector Machine (FSVM), and Neural Network (NN) classifier. The study showed that hybrid sampling technique with CMTFSVM followed by SMOTE using FSVM classifier provided better result on imbalanced dataset.
First stage of hybrid sampling technique was undersampling using CMTFSVM for the majority class. CMTFSVM used truth model and falsity model. Truth model is the real instance of dataset. Falsity model is similar to truth model, however in the falsity model the target output is the complement of the truth model output. The truth model was trained by truth memberships, on the other hand the falsity model was trained by false membership. The membership represented the importance of data to their classes. Membership value aims to surpress the effect of noise and outlier in a class. The membership function define as (1a) and (1b). Membership value for minority class represented by \(m^+_i\) with \(x^+_i\) as data example, while \(m^-_i\) represent the membership value of negative class example \(x^-_i\). The within-class importance for each class represent by \(f(x_i)\), which has value between 0 and 1. The ratio of imbalanced class was reflected by \(r^-\) which reflected the negative/majority class ratio and \(r^+\) which reflected positive/minority class ratio where \(r^>r^-\). Membership value is between [0, \(r^-\)] interval, where \(r^-<1\) reflected negative class. In order to reflect the imbalanced class, the fuzzy membership value has to be \(r^-=r^-\) and \(r^+=1\) for training example.

\[
\begin{align*}
\text{\(m^+_i\)} &= f(x^+_i) \times r^+ \\
\text{\(m^-_i\)} &= f(x^-_i) \times r^-
\end{align*}
\]

(1a)

(1b)

\(F(x_i)\) is based on the distance of data to the actual hyperplane defined as (1c). Variable \(\beta\) is the steepness of the decay increased by 0.1 in [0,1] interval, where \(d_{\text{hyp}}\) is a functional margin of each example, which is equivalent with absolute value of the SVM decision value defined as (1d). The variation of \(\beta\) makes ten possible fuzzy membership values. The membership value for each example selected based on the highest value on Gmean.

\[
\begin{align*}
\text{\(d_{\text{hyp}}\)} &= \frac{2}{1+\exp\left(-|y_i(\omega \times \Phi(x_i) + b)\right)} \\
\end{align*}
\]

(1c)

(1d)

There are two types of undersampling technique which are CMTFSVM1 (CMT1) and CMTFSVM2 (CMT2). Training dataset in CMTFSVM1 created by eliminating all misclassification samples by the truth and falsity model, however CMTFSVM2 eliminated samples which appear both in truth and falsity model.
The next step of hybrid sampling technique implementation was balancing dataset after CMTFSVM method using SMOTE. SMOTE replicated the minority data based on nearest neighbor to create synthetic data and made balanced dataset[8]. Synthetic data created larger data with less specific decision area. The more region created by synthetic data the larger minority class area created. The synthetics data generation process could be described as follows:

1. Selected minority data sample and k nearest neighbor
2. Calculated the different (dif) of sample with its neighbor
3. Randomized gap between 0,1
4. Multiplied dif and gap
5. Synthetics data was the addition of selected sample and result from step 4

2.3. Evaluation
The evaluation of model was conducted using 10-fold cross validation method. We used the evaluation measure of Gmean, AUC, and accuracy using FSVM classifier for all dataset. Gmean evaluation is standard technique to measure an imbalance dataset classifier[6]. Gmean is the square root of the product of true positive rate and true negative rate as shown in equation (2). While, AUC is evaluation measure which used trade off between benefit (TP_rate = TP / (TP + FN)) and cost (FP_rate = FP / (FP + TN)). True Positive (TP) and True Negative (TN) are the instance that classified correctly, on the other hand False Positive (FP) and False Negative (FN) are the instance of negative data and positive data that misclassified.

\[
Gmean = \left( \frac{TP}{TP+FN} \times \frac{TN}{TN+FP} \right)^{\frac{1}{2}}
\]  

(2)

The experiment was conducted using specification of development environment is Personal computer Asus A456U, Intel Core i5-7200U 2.5 Ghz processor with 4 GB RAM. We used the package of ‘mlfromscratch’, python implementation of machine learning package for SVM implementation[9] and the package of ‘imblearn’, python implementation of imbalance learning algorithm for SMOTE implementation[10].

3. Results and discussion
3.1. Data
German dataset was downloaded from UCI, while Glass5 dataset and Yeast3 dataset was downloaded from KEEL. Compound descriptor data were crawled from PubChem using compound ID (CID) from Kurnia’s study[4] and available API. Protein descriptor was obtained from protein feature server (Profeat) using amino acid composition (AAC) feature. Profeat used FASTA file from amino acid sequences, where in this study it was collected from Kurnia’s study[4].

Compound descriptor had fourteen variables that described the compound as numerical value. On the other hand, protein descriptor had 20 variables represented the numeric description of protein. The variables were the proportion of each amino acid divided by the total of amino acid in the protein. Example of compound descriptor and protein descriptor are listed in table 2 and table 3.

| Compound ID | 1005 | 1001 | 100004 | 10001388 | 14505 |
|-------------|------|------|---------|-----------|-------|
| Molecular Formula | C3H5O6P | C8H11N | C20H24N2O2 | C25H26O5 | C6H6O2 |
| Molecular Weight | 168.041 | 121.183 | 324.424 | 406.478 | 110.112 |
| H Bond Donor Count | 3 | 1 | 1 | 3 | 0 |
| H Bond Acceptor Count | 6 | 1 | 4 | 5 | 2 |
| Rotatable Bond Count | 3 | 2 | 3 | 5 | 1 |
| Complexity | 201.0 | 65.0 | 600 | 713 | 98.7 |
| Monoiotopic Mass | 167.982 | 121.089 | 324.184 | 406.178 | 110.037 |
Table 3. Protein descriptor.

| Protein ID | A0P1Y2 | A1A4Y4 | B7U540 | C9JR72 | O00116 |
|------------|--------|--------|--------|--------|--------|
| A          | 8.632  | 7.735  | 8.083  | 14.410 | 8.055  |
| C          | 4.211  | 1.105  | 2.309  | 3.057  | 2.128  |
| D          | 2.316  | 3.867  | 6.005  | 5.022  | 5.015  |
| E          | 3.158  | 7.735  | 7.159  | 5.677  | 6.839  |
| F          | 4.842  | 3.867  | 5.774  | 5.022  | 4.407  |
| G          | 6.316  | 5.525  | 6.928  | 8.734  | 8.967  |
| H          | 6.105  | 1.657  | 3.002  | 1.965  | 2.584  |
| I          | 2.105  | 4.420  | 6.467  | 1.310  | 5.927  |
| K          | 8.211  | 5.525  | 3.695  | 0.873  | 5.927  |
| L          | 8.632  | 8.840  | 9.007  | 12.445 | 7.903  |
| M          | 2.316  | 6.077  | 3.233  | 0.873  | 1.824  |
| N          | 4.421  | 6.630  | 3.233  | 1.528  | 4.103  |
| P          | 10.947 | 3.315  | 3.233  | 6.114  | 4.407  |
| Q          | 4.211  | 3.315  | 3.464  | 3.275  | 3.495  |
| R          | 4.000  | 3.315  | 7.159  | 7.642  | 5.775  |
| S          | 6.947  | 8.287  | 6.005  | 4.367  | 6.231  |
| T          | 5.474  | 7.735  | 4.388  | 5.895  | 5.927  |
| V          | 4.842  | 7.182  | 7.159  | 7.205  | 5.927  |
| W          | 0.000  | 1.105  | 1.155  | 1.747  | 1.216  |
| Y          | 2.316  | 2.762  | 2.540  | 2.838  | 3.343  |

Ijah compound data interaction (CPI) was obtained from Kurnia’s study[4]. If protein had interaction with compound, the class of CPI was positive, otherwise negative. Table 4 shown the example of CPI from Kurnia’s study[4] and table 5 summarises the data that used in this study.

Table 4. Compound protein interaction.

| Compound ID | Protein ID | Weight |
|-------------|------------|--------|
| COM00000014 | PRO00000377| 1      |
| COM00000016 | PRO00001875| 1      |
| COM00000020 | PRO00002082| 1      |
| COM00000020 | PRO00001846| 1      |
| COM00000020 | PRO00000615| 1      |
### Table 5. Summary of data obtained.

| Data               | Source    | Description                                          |
|--------------------|-----------|------------------------------------------------------|
| Protein            | Profeat   | 3 335 protein descriptor data                        |
| Compound           | PubChem   | 7 119 compound descriptor data                       |
| CPI                | Ijah      | 3 693 compound protein interaction                   |
| Amino acid sequence| Uniprot   | 3 335 amino acid sequence in *.fasta                 |

#### 3.2. Data Praprocess

All attributes from Yeast3 dataset and Glass5 dataset were numerics. We conducted normalization to both datasets. While, data preprocessing for German dataset was conducted by transforming the nominal attribute using label encoding, scaling the ordinal attributes, and conducting normalization for numeric attributes. Label encoding create binary -1,1 data for nominal attribute. No data categorized as noise and outlier for all dataset. Ijah dataset is CPI matrix which consist compound descriptor, protein descriptor, and CPI information. CPI matrix created from each protein descriptor that crossed with each compound descriptor and labelled using CPI information from Kurnia’s study [4]. There are missing Pubchem CID from list of compound in Kurnia’s study [4], it caused some compound does not have compound descriptor. CPI matrix which does not have compound descriptor are deleted. After the deletion, there are 23 741 865 record data for Ijah dataset which consist 2 908 positive class. The example of Ijah dataset is listed in table 6.

### Table 6. Ijah dataset.

| Properties         | Data-0 | Data-1 | Data-2 | Data-3 | Data-14574 |
|--------------------|--------|--------|--------|--------|------------|
| A                  | 0.00634| 0.00837| 0.00754| 0.00587| 0.0005     |
| C                  | 0.00078| 0.00232| 0.00128| 0.00181| 0.00024    |
| D                  | 0.0057 | 0.00604| 0.00384| 0.00158| 0.00024    |
| E                  | 0.00686| 0.00651| 0.00901| 0.00406| 0.00029    |
| F                  | 0.00453| 0.00558| 0.00297| 0.00316| 0.00041    |
| G                  | 0.00621| 0.00372| 0.00544| 0.00429| 0.00036    |
| H                  | 0.00311| 0     | 0.00242| 0.00135| 0.00021    |
| I                  | 0.00531| 0.00465| 0.00416| 0.00565| 0.00036    |
| K                  | 0.00686| 0.0079 | 0.00768| 0.00497| 0.00018    |
| L                  | 0.00958| 0.00697| 0.01042| 0.00948| 0.0008     |
| M                  | 0.00272| 0.00186| 0.0016  | 0.00158| 9.00E-05   |
| N                  | 0.00272| 0.00604| 0.00288| 0.00271| 0.00014    |
| P                  | 0.00492| 0.00325| 0.00585| 0.00181| 0.00039    |
| Q                  | 0.00401| 0.00511| 0.00553| 0.00203| 0.00012    |
| R                  | 0.00479| 0.00465| 0.00658| 0.00226| 0.00027    |
| S                  | 0.00583| 0.00558| 0.0074  | 0.00452| 0.00052    |
| T                  | 0.00531| 0.00604| 0.00475| 0.00429| 0.00046    |
| V                  | 0.00608| 0.00744| 0.00553| 0.00497| 0.00042    |
| W                  | 0.00104| 0.00139| 0.00073| 0.0009 | 0.00026    |
| Y                  | 0.00311| 0.00279| 0.00297| 0.00271| 0.00014    |
| MolecularWeight    | 0.41919| 0.40445| 0.42632| 0.42612| 0.32879    |
This study used 14,575 data from Ijah dataset that classified as negative class (no interaction between protein and compound) and 2,908 data from positive class (compound and protein has interaction). Sample data which used 17,483 record with positive class proportion was 20%. After generating the sample data, all features were normalized using sklearn library.

3.3. Evaluation
The Gmean and AUC resulted in this study were compared to those of Pruengkarn[6] as described in table 7 dan table 8. This comparison showed that the results of this study were differ to those of previous research. It probably was caused by the different techniques used in the preprocessing. In addition, in Pruengkarn[6], in the oversampling using SMOTE, the variables used in their experiments were not explained clearly. In this study, the variable of nearest neighbor (k) was set into equal or greater than 2 (k ≥ 2).

Table 7 shows the evaluation of undersampling, oversampling and hybrid sampling approach. CMT1 and CMT2 was represent two type of undersampling technique which are CMTFSVM1 and CMTFSVM2 respectively, SMOTE was oversampling approach, however CMTSMT1 and CMTSMT2 was hybrid sampling approach used two types of undersampling technique combined by oversampling technique, SMOTE (SMT).

| Dataset      | Approach | Gmean of this study | Gmean of Pruengkarn | ΔGmean  |
|--------------|----------|---------------------|---------------------|---------|
| German       | CMT1     | 0.5828              | 0.7244              | -0.1416 |
|              | CMT2     | 0.5828              | 0.7262              | -0.1792 |
|              | SMOTE    | 0.6434              | 0.7294              | -0.0860 |
|              | CMTSMT1  | 0.6389              | 0.7141              | -0.0752 |
|              | CMTSMT2  | 0.6389              | 0.7367              | -0.0978 |
| Glass5       | CMT1     | 0.7155              | 0.0000              | +0.7155 |
|              | CMT2     | 0.7155              | 0.1000              | +0.7155 |
|              | SMOTE    | 0.8209              | 0.9625              | -0.1416 |
The implementation of hybrid sampling technique measured by Gmean and AUC showed slightly different value. It meant that the implementation of hybrid sampling technique in this study was quite similar to the study conducted by Pruengkarn [6]. The implementation of hybrid sampling technique then was applied to active compound and protein interaction in Ijah dataset. In order to know the effect of hybrid sampling technique in Ijah dataset, the unbalanced dataset was classified using SVM. Table 9 showed how the performance of SVM classifying the unbalanced dataset. The classification performance had 0.0000, 0.5000, and 0.7677 for Gmean, AUC, and accuracy, respectively. The result showed that SVM classifier was failed to identify the minority class.

Table 9. Unbalance Ijah dataset classified using SVM.

| Measurement   | Gmean | AUC  | Accuracy |
|---------------|-------|------|----------|
| Minority samples | Total of minority samples | 2 908 |
| Minority : majority proportion | 0.2 : 0.8 |

Table 10 presents the result of hybrid sampling technique on Ijah dataset. Gmean and AUC increased up to 0.6812 and 0.0319, respectively. Table 10 showed better accuracy compared to those of Table 9 where data was imbalanced. The result showed that all classifiers could recognise both minority and majority samples. After hybrid sampling technique was implemented, the total instances were 13 116.
data. Comparing the result to the original CPI matrix, filtered 9,851 data which was similar to original CPI matrix.

**Table 10.** Hybrid sampling technique on Ijah dataset.

| Measurement       | Gmean    | AUC      | Accuracy |
|-------------------|----------|----------|----------|
| Minority samples  | Total of minority samples | 9,851    | 0.4 : 0.6 |

4. Conclusion
This study used the combination of Complementary Fuzzy Support Vector Machine and Synthetic Minority Oversampling Technique as hybrid sampling technique to handle imbalanced data. The implementation had Gmean which was differ from -0.0752 to -0.1792. Moreover, the value of AUC also was slight difference of -0.0732 to 0.0977 compared to the the previous research conducted by Pruengkarn. The hybrid sampling technique can handle Ijah active compound and protein interaction imbalanced problem. The result showed that the data in minority class was increased up to 150% (9,581 data) from the original data that only consist of 2,908 data. Furthermore, it presented good classification result with 0.8346 and 0.5319 for Gmean and AUC, respectively.

References
[1] Torri M C 2013 European J. Med. Plants 3(1) 25-39
[2] WHO 2008 Traditional medicine fact sheet 134
[3] Amir F 2016 Prediksi formula jamu untuk penyembuhan penyakit diabetes dengan teknik graph mining[thesis] Institut Pertanian Bogor, Bogor, Indonesia
[4] Kurnia A 2017 Prediksi formula jamu berkhasiat menggunakan teknik link prediction dari jejaring bipartite senyawa aktif dan protein [Thesis] (Bogor: Bogor Agricultural University)
[5] Ali A, Shamsuddin S M and Ralescu A L 2015 Int. J. Adv. Soft Comput. Appl. 7(3) 176-204
[6] Pruengkarn R, Wong K W and Fung C C 2017 IEEE Int. Conf. Sys. Man Cybern. (Banff)(Canada) 978-83
[7] Zhang P, Tao L, Zeng X, Qin C, Chen S Y, Zhu F, Yang S Y, Li Z R, Chen W O and Chen Y Z 2016 J. Molecular Biology 429(3) 416-25
[8] Chawla N V, Bowyer K W, Hall L O, and Kegelmeyer W P 2002 J. Artif. Intell. Res. 16 321-57
[9] Noren E L 2017 ML-From-Scratch Github Repository Available at https://github.com/eriklindernoren/ML-From-Scratch
[10] Nogueira F and Aridas C K 2017 J. Mach. Learn. Res. 18(17) 1-5

Acknowledgements
This research is supported by Bogor Agricultural University and Ministry of Research, Technology and Higher Education, Indonesia, under Competitive Research Grant from Directorate of Higher Education, Indonesia, 2017