International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 4, August 2020, pp. 4331~4339
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i4.pp4331-4339

Feature selection for multiple water quality status: integrated bootstrapping and SMOTE approach in imbalance classes

Shofwatul Uyun[1], Eka Sulistyowati[2]
[1]Department of Informatics, Universitas Islam Negeri Sunan Kalijaga, Indonnesia
[2]Department of Biology, Universitas Islam Negeri Sunan Kalijaga, Indonesia

Article Info

Article history:

Received Nov 24, 2019
Revised Feb 17, 2020
Accepted Feb 25, 2020

Keywords:

Bootstrapping
Feature selection
Imbalance class
SMOTE
Water quality status

ABSTRACT

STORET is one method to determine the river water quality, and to classify them into four classes (very good, good, medium and bad) based on the data of water for each attribute or feature. The success of the formation of pattern recognition model much depends on the quality of data. There are two issues as the concern of this research as follows, the data having disproportionate amount among the classes (imbalance class) and the finding of noise on its attribute. Therefore, this research integrates the SMOTE Technique and bootstrapping to handle the problem of imbalance class. While an experiment is conducted to eliminate the noise on the attribute by using some feature selection algorithms with filter approach (information gain, rule, derivation, correlation and chi square). This research has some stages as follows: data understanding, pre-processing, imbalance class, feature selection, classification and performance evaluation. Based on the result of testing using 10-fold cross validation, it shows that the use of the SMOTE-bootstrapping technique is able to increase the accuracy from 83.3% to be 98.8%. While the process of noise elimination onthe data attribute is also able to increase the accuracy to be 99.5% (the use of feature subset produced by the information gain algorithm and the decision tree classification algorithm).

Corresponding Author:

Shofwatul Uyun,
Department of Informatics,
Faculty of Science and Technology,
Universitas Islam Negeri Sunan Kalijaga,
Marsda Adisucipto Street, No. 1 Depok Sleman Yogyakarta, 55281, Indonesia.
Email: Shofwatul.uyun@uin-suka.ac.id

1. INTRODUCTION

STORET is a method used by the Minister of Environment in to determinine water quality status in river/water body [1]. The performance process of STORET method is comparing the data resulting from water sampling with the water quality standard in accordance with the classes and based on the attributes used. The more parameters used may incur more cost related to laboratory handling and measurements. It is because the observation and analysis are conducted in the laboratory for each sample of water data for each sampling point. The number of data analyzed requires automation in determining the water quality status. It requires a model improvement in the pattern recognition field that can be used to classify the water quality status. Generally there are some methods that can be used to measure the water quality status as follows: (a) water quality index as conducted by [2] who suggests The West Java Water Quality Index (WJWQI) to measure the water quality in West Java province, and [3, 4]; (b) based on community suggested by [5]; (c) Water pollution Index [6]; (d) STORET index [7].

Water has a lot of parameters that can be measured to determine its quality status. Based on the value of some selected attributes, the quality status can be classified. In pattern recognition, one of the important components determining the success grade of classification process is the suitable feature

use [8]. There are two process related to feature those are feature extraction process [8] and feature selection process [9]. There are some reasons why feature selection process becomes very important in the pattern recognition as follows: to improve the performance of a model of the pattern recognition system (simple model that has quick performance by eliminating the irrelevant data) [10], to visualize the data on the selection process of model, to decrease the dimension and noise on the data [11]. There are two important issues required to be concerned after the feature extraction process those are the data finding which amount is imbalance among its classes and the noise on the data attribute.

Two approaches are deliberately used to handle the imbalance class case those are for oversampling and under-sampling cases. One technique that can be used to handle both cases is called SMOTE technique [12]. For oversampling case, the duplication of data will be conducted on the minority class. On the other hand, for under-sampling case, some data samples will be eliminated from the majority class or by combining both or usually called the hybrid technique [10]. The use of this technique has the same aim, which is to find the dataset for the learning process having the same data or having almost the same data among the classes (balance). The SMOTE technique has been used to solve the imbalance class case on several studies, among others are the data for detecting the attack [13], the medical data [14-16] and the e-commerce data [17]. Besides SMOTE, there is another technique called bootstrapping that can be used for resampling data. Resampling technique can be used to handle the problem of the data amount on the smaller class from its quantity by changing the distribution of minority class underrepresented during the data training process in the machine learning algorithm. Resampling technique is also known as the solution on imbalance class case for learning dataset. This method is suitable to be used on the data in great scale, which is conducted to decrease the amount of data training sample. So that the training need can use fewer amount of data that represent the actual data.

The noise existence on the attribute data certainly will give impact on its classification performance. If the data used has the very great amount of attribute/parameter or feature, it certainly will give impact during the computing process [11]. Therefore, the feature selection process is required. Generally there are three approaches that can be conducted to select the attribute or feature; including filter approach [18], wrapper approach [19] or embedded approach [19]. In filter approach the process between feature selection and learning is conducted in series. It is different from the wrapper approach that is conducted in parallel. In filter approach, the process of selecting the feature subset is previously conducted based on the weight of each attribute or feature. The weighing is conducted for each attribute or feature to rank the attribute based on the threshold value that has been determined [18].

The classification stage is conducted after obtaining the selected feature. There are several algorithms for the learning process which aim for classification as follows: Decision tree (DT) [18], naive bayes [17], K-nearest neighbors (KNN) [20], random forest [21], artificial neural network [22] and support vector machine [23]. Naïve Bayes is a simple classification model and its learning process does not require a long time if compared with other classification models. Besides, it is also recognized as having good prediction accuracy performance. The use of naïve bayes algorithm is easy and comfortable because it does not need to conduct the complicated parameter estimation and it is reliable to use on the great data [24]. DT is one of the classification algorithms much implemented in several cases of machine learning. The aim of DT is to make a model that can be used to predict the value of a target class on the invisible instance test based on several input features [17, 25]. Some advantages of DT rely on its simplicity, easy to understand, easy to implement, requiring a little knowledge, being able to use in dataset either numeric or categorical, and being able to handle dataset in great amount [26, 27].

Based on the research conducted previously, there is no model integrating the use of bootstrapping resampling technique and SMOTE technique to handle the imbalance class case in multi class case. Besides, the feature selection process by filter approach is conducted to handle the noise on data attribute. There are five algorithms (information gain, rule, chi square, correlation and derivation) used based on the value weight produced and afterwards the performance will be compared among each other. While for classification there are four algorithms (Decision tree, K-nearest neighbours, naïve bayes and random forest).

2. RESEARCH METHOD

This research uses the primary data for one year in Brantas River from November 2017 to October 2018 period. There are 10 locations of sampling which data is analyzed in the Laboratory of Environment Malang as follows: Pendem Bridge, Bumiayu Bridge, Sengguruh Reservoir, Lodoyo Reservoir, Mrican Dam, Ploso Bridge, Lengkong Baru Dam, Porong Bridghe, Gunungsari and Ngujang Bridge. There are twenty two parameters being measured such as temperature, acidity (pH), electrical conductivity (DHL), dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total suspended solid (TSS), total dissolved solids (TDS), Nitrate Nitrogen (NO_3N), Nitrite (NO_2N), PO_4P, H_2S, Phenol, detergent,

free chlorine, oil and fat, Cd, Zn, Cu, Pb, total coliform and Faecal Coliform. The total data used is 120 data samples with 22 parameters. This research has six stages as follows: data understanding, pre-processing, imbalance class, feature selection, classification and performance evaluation, as shown in Figure 1.

The collected dataset has a dimension of 120 rows and 22 columns, the row shows the data taken for each location of taking the river water sample, while column shows the attribute/feature/parameter of water used to determine the status of river water quality. The STORET method is used to determine the status of river water quality based on [1]. STORET method is used to determine the status of water quality. This method compares the data from field measurement with the water quality standard in accordance with waterclass. Hence, Brantas River is included in class 2 category. Before conducting the process of determining the status of water quality, it should previously conduct the following:

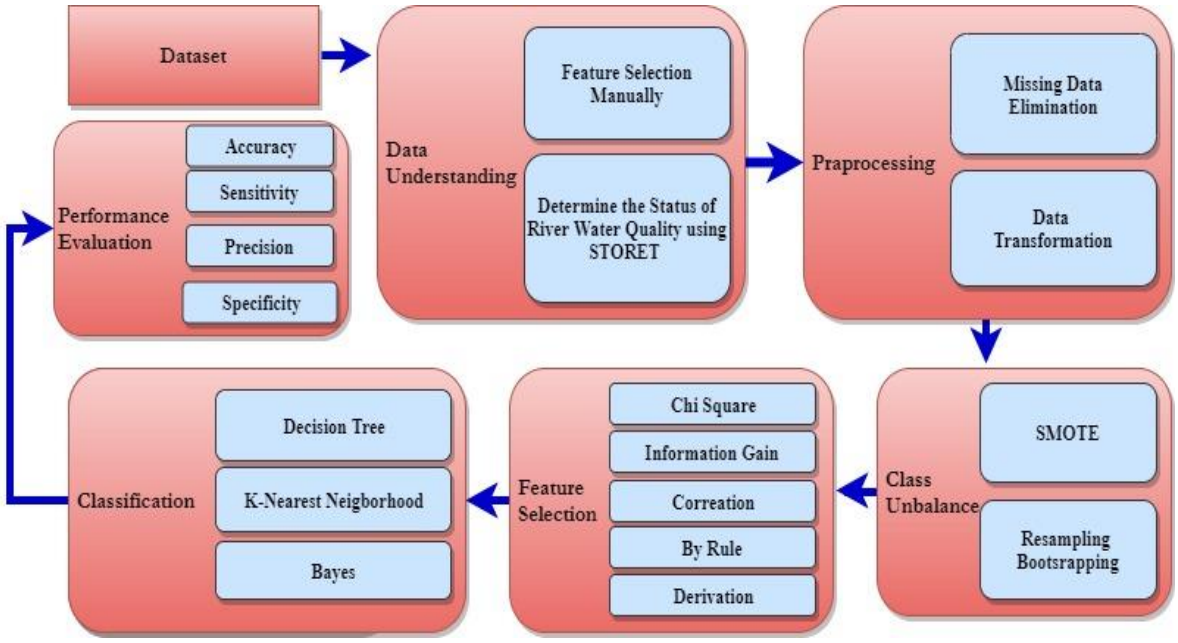

Figure 1. Research stage

2.1. Data understanding

2.1.1. Manual feature sorting

Based on the data result collected, there are some data that cannot be filled completely for all features. It is due to several causes, one of which is each feature/attribute that is not detected by the measuring tool because each has the value under or over the threshold of the measure tool. Based on 22 features, 13 features are selected, while 8 others are not used to determine the status of river water quality.

2.1.2. Determination of status of water quality of Brantas River using STORET method

The determination of status class of the river water quality is conducted based on thirteen selected features. In this case, there are four classes of river quality water as follows: A (excellent), B (good), C (intermediate) and D (bad). The performance process of STORET method is by comparing the data of result of taking the water sample with water quality raw in accordance with its class and based on the parameters used. In this case, Brantas River is included in the second class category for its quality raw. Based on the classification result of the status of river water quality, the unbalance class case is found with the details of classes as follows: A=10, B=16, C=80 and D=14.

2.2. Preprocessing

2.2.1. Missing data elimination

Before conducting the process of selecting the best feature, the process of zero/empty data elimination should be conducted in order not to disturb the performance of algorithm that will be applied to the next process. There are several ways to fill the empty data. It can be filled by the average/minimal/maximal value of data on the feature, or it can be filled with zero value. This research chooses the data average value of the feature.

2.2.2. Data transformation

The data that will be processed needs to be statistically normal to keep staying in one range of the same value. There are several formulations or ways to normalize the data. This research uses proportion transformation. Normalization aims at getting the value on each attribute proportionally.

2.3. Imbalance class

2.3.1. SMOTE (synthetic minority over-sampling technique)

From the result of determining the status of Brantas River water quality using STORET method, it finds a case of unbalance class, so it needs to conduct SMOTE technique [12]. There are two approaches that can be conducted to take SMOTE, with random over-sampling (ROS) and random under-sampling (RUS). Considering the data used for searching the best model to determine the status of river water quality is not too big, ROS approach is selected then. The river water data included in category A, B and D is very minimal so it needs to add the synthetic data taken randomly from the same feature to get the same data mount between the minority and the majority class.

2.3.2. Bootstrapping sampling method

After the dataset is taken from using SMOTE exactly using random over-sampling technique, afterwards it needs to select the data sample on the data training randomly so that the data used has smaller measure.

2.4. Feature selection

The aim of the feature selection process is to eliminate the feature not having a strong contribution in determining the status of water quality. This certainly gives impact on the measure of data dimension either for data training or data testing. Generally there are four approaches to select the feature subset, among others are: filters, which is the process of feature evaluation conducted independently from the learning process; wrappers, which is the process of feature subset selection based on the evaluation result of the learning process; embedded, which is the feature selection conducted during the learning process; and simple filters by assuming the independent feature (this approach is usually used on data with many features such as on the case of textual classification). In this research, the process of feature selection uses filters approach that separates the process of evaluating the best feature subset and the learning process. The determination of the best subset is based on the score or weight produced by each feature subset. The stage of filter approach is shown by Figure 2. This research uses four algorithms included in filters approach category to get the weight value as follows: derivation, information gain, chi square, rule and correlation.

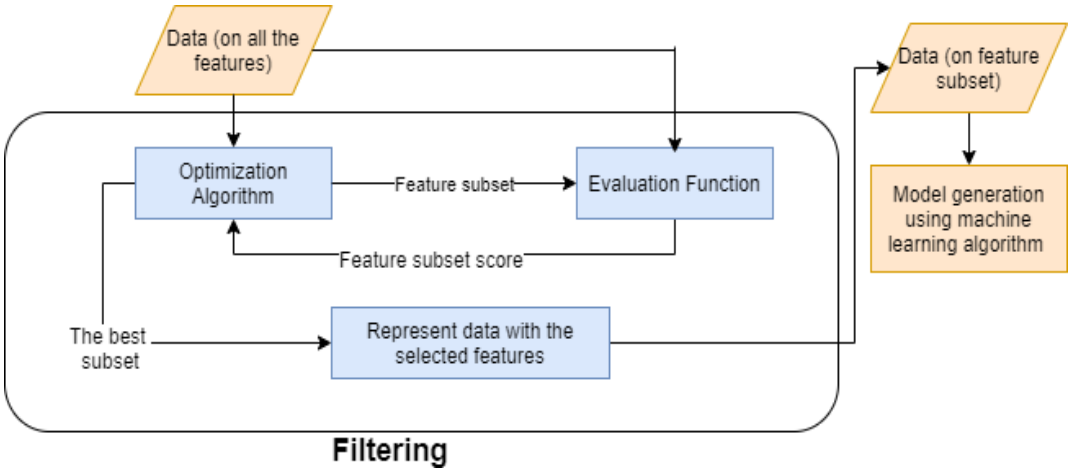

Figure 2. The stage of process of feature subset finding using filters approach

2.5. Classification

The classification stage has a role to find out how far this classification model is able to determine the status of river water quality in the right way based on the data of river for each attribute. There are four algorithms used in this research as follows: Decision tree (DT) [18], naive bayes [17], K-nearest neighbors (KNN) [20] and random forest [21].

2.6. Performance evaluation

The testing process of this research uses k-fold cross validation in distributing the dataset into two parts those are data for training and data for testing alternately for ten times. While several parameters are used to compare the performance between one model and other models those are accuracy, precision, and recall [28].

3. RESULTS OF RESEARCH

3.1. Data understanding

Based on the result of feature selection in manual way and the classification of status of the river water quality, the STORET method is used based on the selected feature. There are four classes those are A, B, C and D. The example of data used in this research is shown in Table 1. There are thirteen selected features (temperature, pH, DHL, DO, BOD, COD, TSS, NO_3N, NO_2N, PO_4P, detergent, total coliform and faecal coliform) based on the unit of quality raw. For example, the data in the fourth column is the data sample of river water in one point of observation with the value for each feature amounted eleven features in category of status of quality A. The values of the total coliform and faecal coliform features are not detected.

Table 1. Data understanding: Data sample with classification result of quality status with thirteen features

No	Description/Parameter	Unit of Quality Standard	1	2	3	4	5	6
	Status Mutu Air (Water Quality Status)		A	C	C	B	C	D
1	Temperatur	mg/l	32.1	29,9	30,2	31	28,1	28,5
2	pH	C	7,29	6,05	7,16	7,18	6,69	7,14
3	DHL	mhos/cm	491	498	472	472	458	514
4	DO		3,8	4,7	4,7	4,2	5	4,8
5	BOD	mg/l	2,6	2,5	3,77	3,69	6,77	4,93
6	COD	mg/l	11,8	19,11	18	23,02	24,58	29,38
7	TSS	Jml/100 ml	40	40	81,4	64	212	206
8	NO3N		2,08	2.263	2.799	2,95	2,4	2,809
9	NO2N		0,187	0,059	0,151	0,123	0,133	0,108
10	Po4P		0,082	0,117	0,087	0,068	0,063	0,087
11	Detergen		0,056	0,012	0,006	0,051	0,098	0,02
12	Total Coliform				430			230
13	Faecal Coliform				230			90

3.2. Pre-processing

In this stage, there are several steps conducted to prepare the dataset that is free from the empty data and to normalize the data to get the good result of classification process. In this case, some experiments of pre-processing method are conducted with the classification method of decision tree using five-fold cross validation with stratified sampling. The result of testing experiment is different from the t-Test to get the best pre-processing method, which is shown in Table 2. Column B shows the accuracy result of using the data that previously has not conducted the normalisation of 79.2%. Column C shows the accuracy result of using the data that previously has conducted the replace process towards the missing value of 82.5%. Column D shows the accuracy result of using the data that previously has conducted the normalization of 79.2 %, and column E shows the accuracy result of using the data that previously has conducted the normalization process and the replace process towards the data with missing value of 83.3%. Based on the difference of the testing result with the t-Test, it shows that the pre-processing process (conducting the normalization of data and the replace towards the missing value) is able to increase the performance of classification result.

Table 2. The testing result is different from the t-Test to get the best pre-processing method

A	B	C	D	E
	0.792 +/- 0.029	0.825 +/- 0.073	0.792 +/- 0.059	0.833 +/- 0.088
0.792 +/- 0.029		0.351	1.000	0.328
0.825 +/- 0.073			0.276	0.820
0.792 +/- 0.059				0.229
0.833 +/- 0.088				

3.3. Imbalance class

The dataset obtained in data understanding stage has imbalance data in each class. This condition later will give effect on the data training process. Therefore, three scenarios are conducted at this stage as follows: SMOTE, bootstrapping and integration between SMOTE and bootstrapping, in which the training and testing process are conducted with the decision tree method using 10 fold-cross validation. The different

testing experiment result with the t-Test in handling the imbalance class case is shown in Table 3. Column B shows the use of SMOTE method, column C shows the use of bootstrapping method, and column D is the integration between SMOTE method and bootstrapping. Based on the explanation of Table 3, it shows that the use of SMOTE method is able to increase the accuracy result of training of 96.5% and the accuracy result of training process keeps increasing using the integration method between SMOTE and bootstrapping of 98.8%.

Table 3. Different testing result with t-Test to get the method to handle the imbalance class case

A	B	C	D
	0.965 +/- 0.025	0.858 +/- 0.040	0.988 +/- 0.032
0.965 +/- 0.025		0.000	0.089
0.858 +/- 0.040			0.000
0.988 +/- 0.032			

3.4. Feature selection

The target of this stage is to get the best feature in determining the status of river water quality. There are five algorithms used in conducting the feature selection with filter approach as follows: information gain, chi square, derivation, correlation and by rule. The coding is conducted previously for each feature. F_1=Temperature; F_2 =pH; F_3=DHL; F_4 =DO; F_5=BOD; F_6=COD; F_7=TSS; F_8=NO_3N; F_9=NO_2N; F_{10} =P_o4P; F_{11}=Detergent; F_{12}=Total Coliform and F_{13}=Faecal Coliform. The result of selected attribute and feature for each algorithm of feature selection is shown in Table 4.

Table 4. Selected feature set based on several feature selection algorithms

No	Feature Selection Algorithms	Subset Fitur
1	Rule	{F_5, F_7, F_8, F_9, F_3, F_6, F_4, F_{12}}
2	Chi Square	{F_2, F_6, F_4, F_5, F_8, F_{13}, F_1, F_{12}}
3	Information Gain	{F_5, F_{13}, F_{12}, F_6, F_3, F_{11}, F_7, F_8, F_9, F_1}
4	Correlation	{F_2, F_{13}, F_{12}, F_6, F_{11}, F_5, F_3, F_8}
5	Derivation	{F_{10}, F_{11}, F_7, F_9}

Based on the data from Table 4, it can be seen that the finding of feature subset has the best score using five feature selection algorithms with filter approach. For example, in number 1 the second row there are eight feature subsets produced by the algorithm rule those are: BOD, TSS, NO_3N, NO_2N, DHL, COD, DO and Total Coliform. Afterwards, the feature selection result uses chi square algorithm (pH, COD, DO, BOD, NO_3N, Faecal Coliform, Temperature, Total Coliform), information gain (BOD, Faecal Coliform, Total Coliform, COD, DHL, detergent, TSS, NO_3N, NO_2N, temperature), correlation (pH, faecal coliform, total coliform, COD, detergent, DHL, DHL, NO_3N) and derivation (P_o4P, detergent, TSS, NO_2N), which is shown in Table 4 in the next row with several selected feature subsets. Afterwards a learning process is conducted from those several feature subsets using the decision tree method to know the performance and the result is shown in Table 5. Column B to column F show the classification result using the selected feature subset produced using several feature selection algorithms (chi square, derivation, information gain, correlation and rule). The t-Test testing result shows that the use of selected feature subset produced by the information gain algorithm has the highest accuracy value of 99.5%.

3.5. Classification

After the best feature subset has been obtained, which is produced by some algorithms, a classification is conducted using four classification algorithms then the average value is calculated from the use of the selected feature subset. The classification algorithm used are: decision tree, k-NN, naïve bayes and random forest. Based on the data shown in Table 6 and Figure 3, it can be found out that the result of classification using eight feature subsets produced by chi square algorithm with the highest accuracy value is produced by the decision tree algorithm of 98.50% with the accuracy average for four classification algorithms of 96.29%. While the result of classification using four feature subsets produced by the derivation algorithm with the highest accuracy value is produced by the random forest algorithm of 98.49% with the accuracy average for four classification algorithms of 91.86%. The use of feature subset produced by the information gain algorithm amounted ten feature subsets is able to produce the highest accuracy value with the decision tree and random forest classification algorithms of 99.50%. While for the average of 96.92% the different result is also shown by the rest of the two algorithms those are correlation and rule.

Both produce the best accuracy value by the same classification algorithm that is random forest of 97.99% and 99.50%. Generally, it can be concluded that feature subset produced by the information gain and random algorithms is able to produce the accuracy level more than 96.5%.

Table 5. Result of T-test towards classification result using decision tree algorithm and selected feature subset

A	B	C	D	E	F
	0.985 +/- 0.017	0.950 +/- 0.047	0.955 +/- 0.016	0.977 +/- 0.032	0.988 +/- 0.032
0.985 +/- 0.017		0.041	0.196	0.520	0.830
0.950 +/- 0.047			0.010	0.146	0.051
0.995 +/- 0.016				0.138	0.512
0.977 +/- 0.032					0.488
0.988 +/- 0.032					

Table 6. Classification result uses feature subset produced by feature selection process

	Classification Algorithm	Accuracy	Recall	Precision
Chi Square	Decision Tree	98,50%	98,47%	98,70%
	k-NN	96,73%	96,74%	97,15%
	Naïve Bayes	92,22%	92,37%	92,75%
	Random Forest	97,74%	97,72%	97,97%
Derrivation	Decision Tree	94,99%	95,25%	95,53%
	k-NN	95.48%	95.49%	95.99%
	Naïve Bayes	78.38%	79.18%	81.28%
	Random Forest	98.49%	98.55%	98.65%
Information Gain	Decision Tree	99,50%	99,50%	99,50%
	k-NN	97.22%	97.22%	97.75%
	Naïve Bayes	91.46%	91.57%	92.16%
	Random Forest	99.50%	99.50%	99.50%
Correlation	Decision Tree	97,74%	97,77%	98,10%
	k-NN	97.24%	97.22%	97.73%
	Naïve Bayes	91.97%	92.15%	92.92%
	Random Forest	97.99%	98.02%	98.37%
Rule	Decision Tree	98.75%	98.84%	99.06%
	k-NN	98.24%	98.32%	98.29%
	Naïve Bayes	94.98%	95.23%	95.48%
	Random Forest	99.50%	99.52%	99.55%

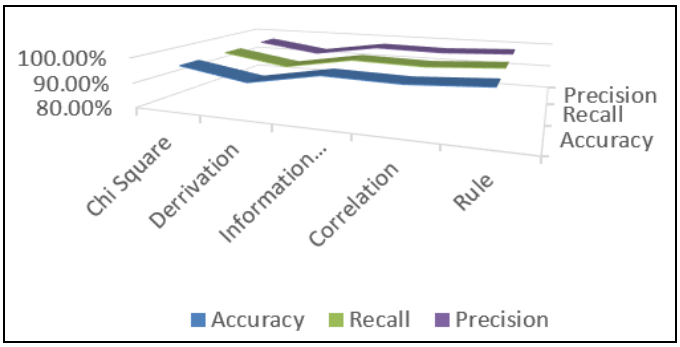

Figure 3. Performance comparison of the use of selected feature based on the average result (accuracy, recall and precision) using four classification algorithms

3.6. Performance evaluation

Generally the model of pattern recognition for the classification of the status of river water quality based on several water feature subsets has the sub stage of process as follows: without pre-processing, pre-processing, SMOTE technique, and bootstrapping to handle the imbalance class and the feature selection. In this case, a comparison for each sub-process is conducted using the decision tree algorithm in the classification process. Based on the testing result using 10-fold cross validation, the accuracy average value is obtained as seen in Figure 4.

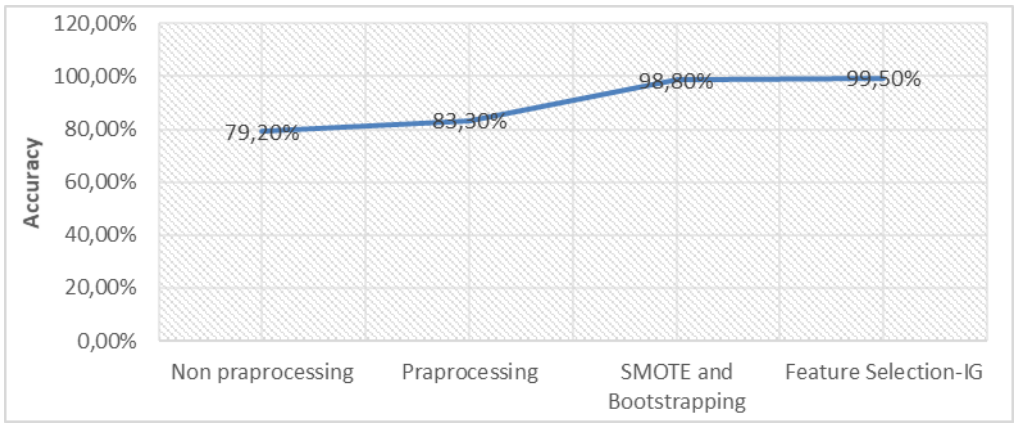

Figure 4. Performance comparison for each step using decision tree algorithm in its classification stage

4. CONCLUSION

The amount of data that is imbalanced in each class is proved to give effect on the learning process on the system of pattern recognition. The SMOTE technique and bootstrapping are proved to be able to handle the imbalance class case, in which there is a significant increase in the accuracy value from 83.3% to 98.8%. While to decrease the noise in the attribute, some experiments have been conducted using five feature selection algorithms (chi square, correlation, derivation, information gain and rule). If seen from the average, the use of feature produced by the rule algorithm and the information gain algorithm has the best accuracy value of 97.87% and 96.92%. The use of selected feature using the information gain with the decision tree classification algorithm shows the increase in the accuracy level of 99.5%.

REFERENCES

[1] Keputusan Menteri Negara Lingkungan Hidup, "Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tentang Pedoman Penentuan Status Mutu Air," *Jakarta Menteri Negara Lingkung Hidup*, pp. 1–15, 2003.
[2] A. D. Sutadian, N. Muttil, A. G. Yilmaz, and B. J. C. Perera, "Development of a water quality index for rivers in West Java Province, Indonesia," *Ecol. Indic*., vol. 85, pp. 966-982, 2018.
[3] M. Bora, and D. C. Goswami, "Water quality assessment in terms of water quality index (WQI): case study of the Kolong River, Assam, India," *Appl. Water Sci*., vol. 7, no. 6, pp. 3125-3135, 2016.
[4] K. A. Shah, and G. S. Joshi, "Evaluation of water quality index for River Sabarmati, Gujarat, India," *Appl. Water Sci*., vol. 7, no. 3, pp. 1349-1358, 2017.
[5] T. Carlson, and A. Cohen, "Linking community-based monitoring to water policy: Perceptions of citizen scientists," *J. Environ. Manage*., vol. 219, pp. 168–177, 2018.
[6] R. L. Kaswanto, H. S. Arifin, and N. Nakagoshi, "Water quality index as a simple indicator for sustainability management of rural landscape in West Java, Indonesia," *Int. J. Environ. Prot.*, vol. 2, no. 12, pp. 17–27, 2012.
[7] R. Y. Tallar, and J. P. Suen, "Identification of waterbody status in Indonesia by using predictive index assessment tool," *Int. Soil Water Conserv. Res*., vol. 3, no. 3, pp. 224-238, 2015.
[8] E. Salahat, and M. Qasaimeh, "Recent advances in features extraction and description algorithms: A comprehensive survey," *Proc. IEEE Int. Conf. Ind. Technol*., pp. 1059-1063, 2017.
[9] S. Uyun, and L. Choridah, "Feature selection mammogram based on breast cancer mining," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 1, pp. 60-69, 2018.
[10] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A novel ensemble method for imbalanced data learning," *Comput. Intell. Neurosci*., pp. 1-11, 2017.
[11] L. Ma, and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1-18, 2017.
[12] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res*., vol. 61, pp. 863-905, 2018.
[13] M. Reza, S. Miri, and R. Javidan, "A hybrid data mining approach for intrusion detection on imbalanced NSL-KDD Dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 6, pp. 20-25, 2016.
[14] B. Krawczyk, M. Galar, Ł. Jeleń, and F. Herrera, "Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy," *Appl. Soft Comput. J.*, vol. 38, pp. 714-726, 2016.
[15] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project," *PLoS One*, vol. 12, no. 7, pp. 1-15, 2017.

[16] N. Mustafa, J.-P. Li, R. A., And M. Z., "A Classification model for imbalanced medical data based on PCA and farther distance based synthetic minority oversampling technique," *Int. J. Adv. Comput. Sci. Appl*., vol. 8, no. 1, pp. 61-67, 2017.
[17] A. Saputra and S. -, "Fraud detection using machine learning in e-Commerce," *Int. J. Adv. Comput. Sci. Appl*., vol. 10, no. 9, pp. 332-339, 2019.
[18] O. Osanaiye, H. Cai, K. K. R. Choo, A. Dehghantanha, Z. Xu, and M. Dlodlo, "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing," *Eurasip J. Wirel. Commun. Netw.*, no. 1, 2016.
[19] A. Mustaqeem, S. M. Anwar, and M. Majid, "Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants," *Computational and Mathematical Methods in Medicine.*, no. 1, pp. 1-10, 2018.
[20] J. Wäldchen, and P. Mäder, "Plant species identification using computer vision techniques: A systematic literature review," *Springer*, vol. 25, no. 2, 2018.
[21] Z. M. Hira, and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinformatics*, no. 1, 2015.
[22] C. Incorvaia, *et al*., "The soft computing-based approach to investigate allergic diseases: A systematic review," *Clin. Mol. Allergy*, vol. 15, no. 1, pp. 1-14, 2017.
[23] R. W. D. Pedro, A. Machado-Lima, and F. L. S. Nunes, "Is mass classification in mammograms a solved problem? -A critical review over the last 20 years," *Expert Syst. Appl*., vol. 119, pp. 90-103, 2019.
[24] K. S. Reddy and E. S. Reddy, "Integrated approach to detect spam in social media networks using hybrid features," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, p. 562, 2019.
[25] T. A. Assegie, and P. S. Nair, "Handwritten digits recognition with decision tree classification: A machine learning approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, pp. 4446-4451, 2019.
[26] J. Singh, K. Singh, and J. Singh, "Reengineering framework for open source software using decision tree approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 3, pp. 2041-2048, 2019.
[27] R. N. Rithesh, R. Vignesh, and M. R. Anala, "Autonomous traffic signal control using decision tree," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 3, pp. 1522-1529, 2018.
[28] I. Sumaiya Thaseen, and C. Aswani Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *J. King Saud Univ.-Comput. Inf. Sci*., vol. 29, no. 4, pp. 462-472, 2017.

BIOGRAPHIES OF AUTHORS

Dr. Shofwatul 'Uyun, S.T., M. Kom is a full time lecturer at the department of Informatics and Head of Information Technology and Database, Universias Islam Negeri (UIN) Sunan Kalijaga in Yogyakarta, Indonesia. She obtained her Bachelor degree in Informatics from Islamic University of Indonesia. She received her M. Kom. and Dr in Computer Science from the Gadjah Mada University. Her research interests are pattern recognition, artificial intelligence and medical image processing.

Eka Sulistiyowati, MA, MIWM is a full time teaching at the Biology Education Study Programme at State Islamic University (UIN) Sunan Kaliajaga Yogyakarta. Environmental management by training, she obtained her degree on Integrated Water Management from The University of Queensland, Australian. Her research interest ranges from environmental management, resource management, water, and biodiversity conservation