Improving DNA Barcode-based Fish Identification System on Imbalanced Data using SMOTE

Wisnu Ananta Kusuma*[1], Nurdevi Noviana[2], LailanSahrina Hasibuan[3], Mala Nurilmala[4]
[1,2,3]Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University, Bogor, Indonesia
[1,3]Working Group of Bioinformatics, Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University, Bogor, Indonesia
[4]Department of Aquatic Product Technology, Faculty of Fisheries and Marine Sciences,
Bogor Agricultural University, Bogor, Indonesia
*Corresponding author, e-mail: ananta@apps.ipb.ac.id

Abstract

Problem in imbalanced data is very common in classification or identification. The problem is raised when the number of instances of one class far exceeds the other. In the previous research, our DNA barcode-based Identification System of Tuna and Mackerel was developed in imbalanced dataset. The number of samples of Tuna and Mackerel were much more than those of other fish samples. Therefore, the accuracy of the classification model was probably still in bias. This research aimed at employing Synthetic Minority Oversampling Technique (SMOTE) to yield balanced dataset. We used k-mers frequencies from DNA barcode sequences as features and Support Vector Machine (SVM) as classification method. In this research we used trinucleotide (3-mers) and tetranucleotide (4-mers). The training dataset was taken from Barcode of Life Database (BOLD). For evaluating the model, we compared the accuracy of model using SMOTE and without SMOTE in order to classify DNA barcode sequences which is taken from Department of Aquatic Product Technology, Bogor Agricultural University. The results showed that the accuracy of the model in the species level using SMOTE was 7% and 13% higher than those of non-SMOTE for trinucleotide (3-mers) and tetranucleotide (4-mers), respectively. It is expected that the use of SMOTE, as one of data balancing technique, could increase the accuracy of DNA barcode based fish classification system, particularly in the species level which is difficult to be identified.

Keywords: *DNA Barcode, imbalanced dataset, mislabeled fish, SMOTE, support vector machine*

1. Introduction

One of the problems in fishery processed product is fish fraud, in which the content of product, especially tuna or mackerel, is replaced by the low price fish. This substitution will harm consumers and can cause a serious problem in health. To minimize this problem, the DNA barcode based identification system could be applied to overcome the limitation of the identification technique based on morphology which is not proper to be implemented to the processed product as conducted by [1, 2].

DNA barcoding is a technique which could provide a biology barcode consisting of short DNA sequences (400-800 bp) standardized to identify a species [3, 4]. Unlike molecular phylogenetic, DNA barcodes are not intended to find the patterns among species but rather to determine the unknown samples [5] and to assess whether the samples should be combined or separated [6]. This idea aimed to distinguish species and to identify specimen such as organ pieces or processed material using short DNA sequence [7]. DNA barcoding use the information of one or several regions in gen. The most common used for DNA Barcode is almost 600 bp segments from *cytochrome oxidase* 1 (CO1) in mitochondria genome (mtDNA) [8]. The other DNA barcode could be obtained from *Cytochrome* b (*cyt* b), a protein found in cell mitochondria of eukaryotic cell. This protein has a role as a part of the electron transport chain and as a subunit of trans-membrane *cytochrome* bc1 and b6f complex [9].

According to Pati [10], the DNA barcode based identification process could be divided into two approaches, namely homology based identification approach and composition based identification approach. The homology based approach is conducted by aligning DNA query

Received May 1, 2017; Revised August 2, 2017; Accepted August 17 2017

fragment to the reference sequence existing in database, such *Barcode of Life Database* (BOLD). Several studies have been conducted using this approach such as [10, 11]. The results of Benedict's research showed a high accuracy of identifying sashimi tuna fillets and cream dory products. However, there was a high probability of incorrect identification of Bluefin Tuna fillet. Moreover, Lowenstein, et al., [12] reported up to 100% accuracy when identifying tuna sushi using a character-based and BLAST.

Unlike the homology based approach, the composition based approach is performed by calculating the frequencies of subsequence existing in DNA fragment and used these frequencies as features as inputs for a machine learning algorithm. This subsequence is commonly named as k-mers. This approach overcomes the drawback of homology approach in term of computational time by avoiding pairwise alignment for each DNA fragments. Several research related to the composition based approach for classifying DNA sequence have been conducted by Seo [13] and Weitschek [14]. Seo [13] used SVM and k-mers frequency to identify the location of a specific pattern on the species. Moreover, Weitschek [14] compared some machine learning method such as Support Vector Machine (SVM), Naïve Bayes, RIPPER, and C4.5 in order to classify DNA sequence. The results showed that SVM could outperform the other machine leaning methods.

The use of SVM, k-mers frequencies of DNA barcode sequence for identifying fish contained in processed product was conducted in our previous work by Mulyati, et al., [15]. This study developed the classification model for identifying Tuna and Mackerel. The accuracy of using tetranucleotide frequency as feature was higher than that of using trinucleotide frequency. The accuracy values are 99.45% and 88% for using tetranucleotide and trinucleotide, respectively. However, the dataset used in this study was still in imbalance. Thus, the accuracy had potentiality to be bias since the class major dominated the decision of classification [16]. The problem of imbalanced dataset can be solved using data balancing technique. Chawla *et al.* [16] employed *Synthetic Minority Oversampling Technique* (SMOTE) for improving the classification model.

This research employed SMOTE [16] for handling the problem faced by imbalanced dataset in the case of fish identification of processed product [15]. We used SVM as classification method and k-mers frequencies of DNA barcoding sequences as features. In this study we used composition based approach used in [13, 14] by chosing trinucleotide (3-mers) and tetranucleotide (4-mers) frequencies as features. SVM is very popular as one of machine learning techniques which has high performance to solve the problem of classification or identification in many cases [17, 18].

2. Research Method

2.1. Dataset

We used DNA barcode sequence from Barcode of Life Database (BOLD) (http://boldsystem.org) as dataset for training. BOLD is an informatics workbench which could help to retrieve, store, analyze, and publish DNA barcode [19]. The data is stored in FASTA format. This dataset consists of 26 species which belong to 7 genus with the total number of 1089 DNA barcode sequence. This dataset belongs to three classes, namely Tuna, Mackerel, and other fish. This dataset actually is still in imbalance. The number of DNA barcode sequences of Tuna and Mackerel are more than other fish. The detail of training dataset is described on the Table 1.

To evaluate the model, we used testing dataset obtained from BOLD and from Department of Aquatic Product Technology, Faculty of Fisheries and Marine Sciences, Bogor Agricultural University, Bogor, Indonesia (Table 2). This dataset also consists of 26 species and 7 genus with the total number of 235 DNA barcode sequence and belong to three classes as those in training dataset.

2.2. Methods

2.2.1. Features Extraction and Normalization

Features extraction of DNA barcode sequences was conducted by counting k-mers frequencies of each sequence. The output of this step is a composition matrix that would be inputted to SVM. We used trinucleotide (3-mers) and tetranucleotide (4-mers). Thus, we had composition matrixes consisted of 64 features and 256 features for 3-mers and 4-mers, respectively. This step was implemented to both training and testing dataset. Thus, we had composition matrixes of training data with the size of 854 x 64 for trinucleotideand 854 x 256 for tetranucleotide. Moreover, the sizes of composition matrixes of testing data were 235 x 64 for trinucleotideand 235 x 256 fortetranucleotide. Next, the composition matrixes were normalized using Equation (1). According to Han, et al., [20], normalization aimed to yield data that have range of value between 0 and 1, but the characteristics of data are not lost.

$$v' = \frac{v - min_a}{max_a - min_a} \quad (1)$$

In the Equation (1), min_a and $max_a$is the minimum value and the maximum value of feature A, respectively. Moreover, $v$is the value of feature A for each sample which is transformed into the range of 0 to 1 using Equation (1).

Table 1. Training dataset of DNA barcode sequence taken from BOLD

Genus	Species	The number of DNA *Barcode sequences*	The average of length of DNA Barcode (Bp)	Class
Thunnus	*alalunga*	119	675	Tuna
	atlanticus	25	777	
	albacares	118	695	
	maccoyi	10	752	
	obesus	88	679	
	orientalis	11	691	
	thynus	131	647	
	tonggol	20	831	
Scomberomorus	*brasiliensis*	12	682	Mackerel
	cavalla	14	745	
	commerson	43	621	
	guttatus	6	892	
	maculatus	12	929	
	munroi	4	746	
	munroi x semifasciatus	4	701	
	niphonius	34	704	
	plurilineatus	6	690	
	queenslandicus	4	702	
	regalis	16	681	
Carcharhinus	*acronotus*	12	632	Other fish
	albimarginatus	24	652	Other fish
Gadus	*macrocephalus*	64	655	Other fish
Hypostomus	*affinis*	10	685	Other fish
	auroguttatus	13	655	Other fish
Lepidocybium	*flavobrunneum*	25	799	Other fish
Lutjanus	*analis*	29	652	Other fish

Table 2. Testing dataset of DNA barcode sequence taken from BOLD dan Department of Aquatic Product Technology, Bogor Agricultural University

Genus	Species	The number of DNA *Barcode sequences*	The average of length of DNA Barcode (Bp)	Class
Thunnus	[1]*alalunga*	9	675	Tuna
	[1]*atlanticus*	3	777	
	[1]*albacares*	52	695	
	[1]*maccoyi*	2	752	
	[1]*obesus*	11	679	
	[1]*orientalis*	1	691	
	[1]*thynus*	63	647	
	[1]*tonggol*	3	831	
Scomberomorus	[1]*brasiliensis*	2	682	Mackerel
	[1]*cavalla*	1	745	
	[2]*commerson*	46	621	
	[1]*guttatus*	1	892	
	[1]*maculatus*	2	929	
	[1]*Munroi*	1	746	
	[1]*munroi x* [1]*semifasciatus*	1	701	
	[1]*niphonius*	2	704	
	[1]*plurilineatus*	1	690	
	[1]*queenslandicus*	1	702	
	[1]*Regalis*	2	681	
Carcharhinus	[1]*acronotus*	2	632	Other fish
	[1]*albimarginatus*	3	652	Other fish
Gadus	[1]*macrocephalus*	15	655	Other fish
Hypostomus	[1]*Affinis*	1	685	Other fish
	[1]*auroguttatus*	2	655	Other fish
Lepidocybium	[1]*flavobrunneum*	3	799	Other fish
Lutjanus	[1]*analis*	5	652	Other fish

[1]Data is taken from BOLD.

[2]Data is obtained from Maulid *et al.* (2016), Department of Aquatic Product Technology, Faculty of Fisheries and Marine Sciences, Bogor Agricultural University

2.2.2. Data Balancing using SMOTE

SMOTE, the data balancing technique employed in this research was first introduced by Chawla, et al., [16] as a technique for solving the problem in imbalanced dataset. This technique was conducted by generating synthetics data. Unlike the common oversampling technique that randomly duplicating data, SMOTE create synthetics data based on k-nearest neighbor that randomly chosen. For numeric data, k-nearest neighbors were measured using Euclidian distance. Firstly, for each attributes, we calculated the difference between minority samples and one of k nearest neighbor value (*i*). This difference was multiplied by random value between 0 and 1. Next, the results were added by the value of minority sample to obtain a new feature vector (a new s*ynthetic minority class* (k_i)).

In this study, we included genus of *Thunnus* and *Scomberomorus* belong to majority class, whereas other fish that consisted of *Carcharhinus, Gadus,* and *Carcharhinus* belong to minority class. The balancing data was conducted by oversampling using SMOTE and under-sampling in a certain proportion. Oversampling was applied to the minority class, while under-sampling was applied to the majority class. The ratio between majority class and minority class of the origin dataset is 677: 177. By applying a combination of under-sampling and oversampling, the initial bias of the learner towards the negative (majority) can be minimized [16].

2.2.3. Classification Method

In this research, the SVM, one of very popular machine learning methods for solving the problem in identification or classification, was chosen to develop a model for DNA barcode sequence identification system. SVM has a good performance for solving many problems of identification [17, 18]. However, the performance of SVM might be decreased when dealing with imbalanced dataset since the number of data in majority class could affect the choosing of the optimal hyperplane. Basically, SVM is used for finding the optimal hyperplane which separated as far as possible two classes of data. The optimal hyperplane could be obtained by maximizing margin, the distance between the optimal hyperplane and sample vectors of training data. Vector of training data located on the margin is named as support vectors. If the training dataset is in imbalance then the choosing of the optimal hyperplane was affected dominantly by samples vectors of majority class, a class which has much more samples data.

Basically, SVM is a binary classification method. In the case of multiclass classification as in this study, we used one-againts-one technique to handling multiclass classification problem. This technique allow us to generate k(k-1)/2 binary classifier models, in which k is the number of class. Next, each sample testing would be classified by all models. The decision was made by voting, with the intention of classifying a sample testing to the class with the most votes.

2.2.4. Experiment Setup

The training data that extracted by k-mers was inputted to the SVM. In the training phase, the optimal value of hyper-parameters C and γ was found using grid search with 10 cross validation in the range of 10^{-6} – 10^{-1} and 10^{-1} – 10^{2} for the parameter of γ and C, respectively. These parameters represented the best model of SVM used in this study (Table 3). The choice of parameters determines the performance of classifiers [21] and the classification results [22].

The SVM was implemented using R programming language which is available on the library e1071 [23]. Kernel function used in this research is Gaussian Radial Basis Function (RBF) recommended by [24]. The best classification model obtained from the training phase was tested using testing dataset downloaded from BOLD and Faculty of Fisheries and Marine Sciences, Bogor Agricultural University, Bogor, Indonesia. Testing aimed at to identify the testing samples into their respective classes.

Table 3. The best parameter value of *γ* and *C*

Features	Taxonomy level	Parameters		Keterangan
		C	γ	
3-mers	genus	100	0.01	Non-SMOTE
3-mers	genus	100	0.0001	using SMOTE
3-mers	spesies	1	0.01	Non-SMOTE
3-mers	spesies	10	0.01	using SMOTE
4-mers	genus	1000	0.001	Non-SMOTE
4-mers	genus	100	0.001	using SMOTE
4-mers	spesies	10	0.001	Non-SMOTE
4-mers	spesies	100	0.1	using SMOTE

2.2.5. Evaluation

The identification results of testing data were represented into confusion matrix and evaluated by several measures such as accuracy and Fmeasure. The Equation (2) to (5) showed the equation of each measure.

$$\textit{Accuracy} = \frac{\sum TP+\sum TN}{\sum TP+\sum TN+\sum FP+\sum FN} \qquad (2)$$

$$\textit{Recall} = \frac{\sum TP}{\sum TP+\sum FN} \qquad (3)$$

$$\textit{Precision} = \frac{\sum TP}{\sum TP+\sum FP} \qquad (4)$$

$$\textit{Fmeasure} = \frac{2\ .\ recall\ .\ prcision}{precision+recall} \qquad (5)$$

The evaluation was also conducted using Basic Local Alignment Search Tool (BLAST) which can be accessed from http://blast.ncbi.nlm.nih.gov/Blast.cgi. Unlike SVM that employs composition based approach, BLAST uses pairwise alignment as the main characteristics of the homology based approach. In this research, BLAST was used for measuring similarity among species of prediction results which were not classified into the right classes.

3. Results and Analysis

3.1. Data Balancing Results

The proportion of each class after conducting balancing data using SMOTE could be seen in Table 4. Data balancing was conducted by under-sampling of majority class and oversampling of minority class. For dataset with features of 3-mers, we conducted under-sampling with 133.5% of original training dataset for genus and species level. Consequently, the ratio between majority and minority class after balancing was changed from 677:177 to 708:708. Moreover, for features of 4-mers, the ratio between majority and minority class of training dataset after balancing become smaller than that of the original one. The comparison of the amount of training data after conducting data balancing using SMOTE and the initial data in genus and species level was described in Table 5 and Table 6.

Table 4. The oversampling dan undersampling on training dataset

features	Taxonomy level	Persentage		Ratio majority and minority class in original data	Ratio majority and minority class after balancing
		Under-sampling	*Over-sampling*		
3-mers	genus	133.5 %	350 %	677 : 177	708 : 708
3-mers	spesies	133.5 %	350 %	677 : 177	708 : 708
4-mers	genus	200 %	100 %	677 : 177	354 : 354
4-mers	spesies	200 %	100 %	677 : 177	354 : 354

Table 5. The comparison of the initial data and after conducting SMOTE in the genus level

Genus	Initial data	After using SMOTE	
		trinucleotide	*tetranucleotide*
Charcarhinus	36	144	146
Gadus	64	259	97
Hypostomus	23	92	36
Lepidocybium	25	97	36
Lutjanus	29	116	39
Scomberomorus	155	177	99
Thunnus	522	531	255

3.2. Classification Results

The classification results were measured using accuracy. Figure 1 showed the accuracy of the classification results using trinucleotide (3-mers) and tetranucleotide (4-mers) as features after conducting data balancing with SMOTE in the genus and species level. The higher taxonomy level yielded the more accurate model. The identification task based on DNA sequences in species level is still very difficult since the species in the same genus or order probably share the similar subsequence (k-mers) in many regions of their DNA sequences. Figure 1 showed that both accuracies of the model using trinucleotide for non-SMOTE and using SMOTE in genus level were over than 90% compared to those in species level which obtain the accuracy of 63% and 81% for non-SMOTE and using SMOTE, respectively. The similar tendency was yielded by the model using tetranuclotide.

In addition, the use of SMOTE for balancing data was only effective to increase the accuracy of the model in species level both for using trinucleotide and tetranucleotide. The accuracy of the model with trinucleotide in species level was increased from 63% to 81% for non-SMOTE and using SMOTE, respectively. However, the difference in accuracy of the model with tetranucleotide was decreased compared to those of using trinucleotide. This showed that the value of k in k-mers frequencies features affects the accuracy of the model in identifying DNA barcode sequence. The higher value of k would produce the longer k-mers or subsequences of DNA barcode sequences. As consequence, the composition of the feature among the DNA sequence of each species or genus would be more different. Thus, the identification process becomes easier. In other word, choosing short value of k in constructing the features of k-mers frequencies is too short to discriminant the DNA sequences among species.

Table 6. The comparison of the initial data and after conducting SMOTE in the species level

Genus	Species	Initial	After using SMOTE	
		data	*trinucleotide*	*tetranucleotide*
Thunnus	*alalunga*	119	129	71
	atlanticus	25	33	14
	albacares	118	112	61
	maccoyi	10	11	7
	obesus	88	101	43
	orientalis	11	15	2
	thynus	131	149	70
	tonggol	20	16	10
Scomberomorus	*brasiliensis*	12	18	7
	cavalla	14	17	8
	commerson	43	35	22
	guttatus	6	4	3
	maculatus	12	15	8
	munroi	4	2	1
	munroi x semifasciatus	4	2	0
	niphonius	34	23	17
	plurilineatus	6	6	3
	queenslandicus	4	3	2
	regalis	16	17	5
Carcharhinus	*acronotus*	12	62	98
	albimarginatus	24	82	37
Gadus	*macrocephalus*	64	258	101
Hypostomus	*affinis*	10	48	16
	auroguttatus	13	44	19
Lepidocybium	*flavobrunneum*	25	98	40
Lutjanus	*analis*	29	116	43

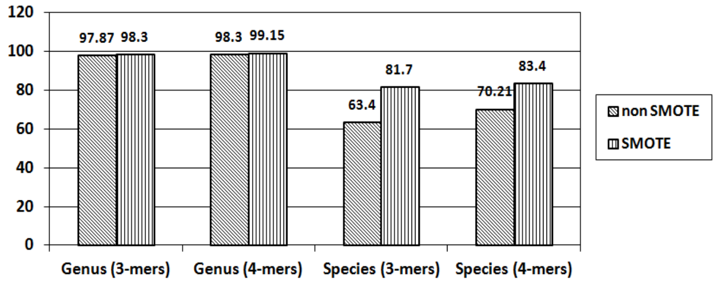

Figure 1. the comparison of accuracy between the model with SMOTE and non-SMOTE using features of trinucleotide (3-mers) and tetranucleotide (4-mers) in genus and species level

Since the dataset is imbalanced, the accuracy of the model is still bias. To evaluate the performance of the model more precisely, we used F-measure as a metric that combines the value of precision and recall as described in Equation (3) to (5) [26]. The average F-measure of the model with trinucleotide in species level are 92% and 94% for non-SMOTE and with SMOTE, respectively. The similar tendency was also shown using tetranucleotide that obtained 92%for non-SMOTE and 95% for using SMOTE. Figure 2 and 3 showed the comparison value of F-measure of non-SMOTE and using SMOTE for trinucleotide and tetranucleotide. Both figures showed that the use of SMOTE could improve the F-measure value almost of all species. It could also be shown that species of *Scomberomorus comerson* and some species of Thunnus had low F-measure. The F-measure of *Scomberomorus commerson* with trinucleotide was 58% and most DNA barcode sequences of *Scomberomorus commerson*was identified as *Thunnus alalunga*and*Thunnus obessus*. The validation using *Basic Local Alignment Search Tool* **(**BLAST**)** onhttp://blast.ncbi.nlm.nih.gov/Blast.cgi showed that *Scomberomorus commerson* has similarity value of 87% with *Thunnus alalunga* and 86% with *Thunnus obessus.*

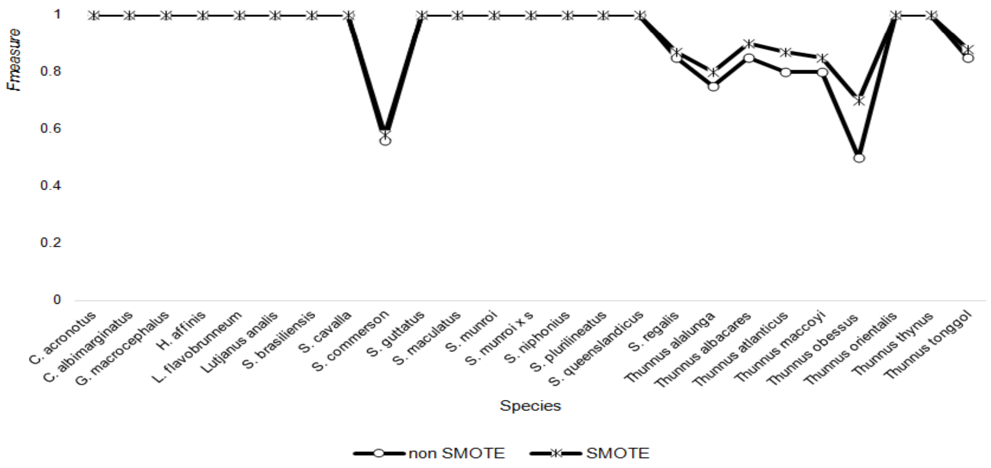

Figure 2. The comparison of F-measure between the model with SMOTE and non-SMOTE using features of trinucleotide

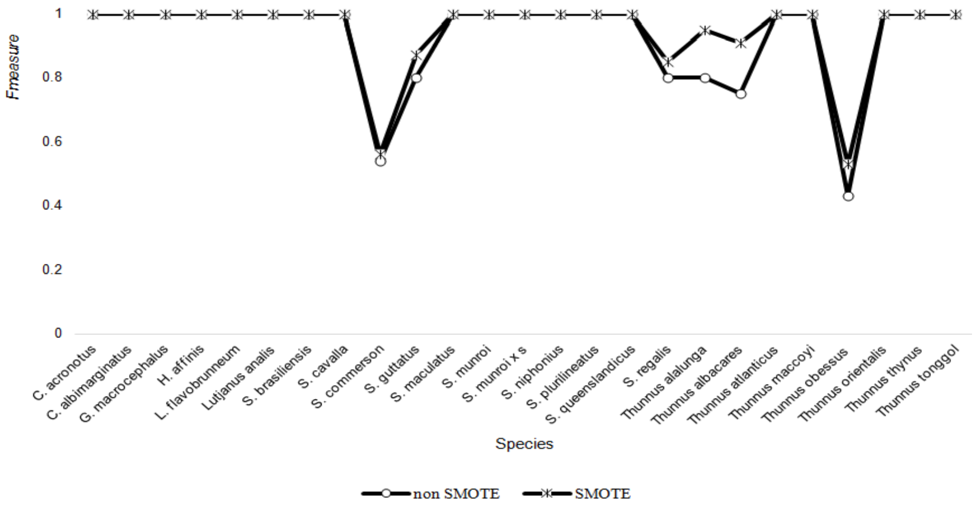

Figure 3. The comparison of F-measure between the model with SMOTE and non-SMOTE using features of tetranucleotide

4. Conclusion

The classification performance of the model representing the DNA barcode based fish identification system could be improved using data balancing with SMOTE in species level. The accuracy of the model using SMOTE was 7% and 13% higher than those of non-SMOTE for trinucleotide (3-mers) and tetranucleotide (4-mers), respectively. However, in the genus level the use of SMOTE only slightly improved the classification performance of the model since the accuracy of the model without SMOTE had been already high, over than 90%. The effect of SMOTE in increasing the classification performance of the model could be seen in the value of F-measure fortrinucleotide and tetranucleotide in the species level. From the F-measure value, we could also notice that the species of *Scomberomorus commerson* had low F-measure. The validation results using BLAST showed that the species *Scomberomorus commerson* had similarity to the species of *Thunnus alalunga* and *Thunnus obessus.* This result was consistent to the fact that many species of *Scomberomorus commerson* was classified to the species of *Thunnus alalunga* and *Thunnus obessus*. In addition, the accuracy of the model would increase by increasing the value of k on k-mers frequencies features.

Acknowledgements

Part of this work was supported by Indonesian Ministry of Research, Technology, and Higher Education through Competency Research Program.

References

[1] Nurilmala M, Widyastuti U, Kusuma WA, Nurjanah, Wulansari N, Widyatuti Y. DNA Barcoding for Identification of Processed Tuna Fish in Indonesian Market. *JurnalTeknologi (Sciences and Engineering).* 2016; 78(4-2): 115-118.

[2] Wulansari N, Nurilmala M, Nurjanah N, Detection Tuna and Processed Products Based Protein and DNA Barcoding. *Indonesian Journal of Aquatic Product Technology*. 2015; 18(2): 120-127.

[3] Hajibabaei, et al. *DNA barcode distinguish species of tropical Lepidoptera*. Proceedings of the National Academic of Sciences. 2009; 103: 968-971.

[4] Hebert PDN, Cywinska A, Ball SL, deWaard JR. *Biological identifications through DNA barcodes*. Proceedings of the Royal Society B: Biological Sciences. 2003; 270(1512): 313-321.

[5] Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. *Use of DNA barcodes to identify flowering plants*. Proceedings of the National Academy of Sciences. 2005; 102(23): 8369-8374.

[6] Koch H. Combining morphology and DNA barcoding resolves the taxonomy of western malagasy liotrigona moure, 1961 (hymenoptera: apidae: meliponini). *African Invertebrates*. 2010; 51(2): 413-421.

[7] Hebert PDN, Ratnasingham S, de Waard JR. *Barcoding animal life: cytochromec oxidase subunit 1 divergences among closely related species*. Proc R Soc. 2003; 270: 96-99.

[8] Seberg O, Petersen G. How many loci does it take to DNA barcode a crocus?. *PLoS ONE*. 2009; 4(2): 4598.

[9] Howell N. Evolutionary conservation of protein regions in the protonmotive cytochromeb and their possible roles in redox catalysis. *J Mol Evol*. 1989; 29(2): 157-169.

[10] Pati A, Heath LS, Kyrpides NC, Ivanova N. ClaMS: A Classifier for Metagenomic Sequences. *Standards in Genomic Sciences.* 2011; 5: 248-253.

[11] Benedict AM, Roselyn DA, Minerva FHV, Sweedy KLP, Mudjekeewis DS. Detection of Mislabeled Commercial Fishery by Products in the Philippines Using DNA Barcodes and its Implications to Food Traceability and Safety. *Food Control*. 2013; 33(1): 119-125.

[12] Lowenstein JH, Amato G, Kolokotronis SO. The Real Maccoyii: Identifying Tuna Sushi with DNA Barcodes-Contrasting Characteristic Attributes and Genetic Distances. *PloS ONE*. 2009; 4(11): 7866.

[13] Seo TK. Classification of Nucleotide Sequences Using Support Vector Machines. *Journal of molecular evolution*. 2010; 71(4): 250-267.

[14] Weitschek E, Fiscon G, Felici G. Supervised DNA Barcodes Species Classification: Analysis, Comparison, and Results. *BMC Bio Data Mining*. 2014.

[15] Mulyati, Kusuma WA, Nurilmala M. Identification of Tuna and Mackerel Based on DNABarcodes using Support Vector Machine. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2016; 14(2): 778-783.

[16] Chawla VN, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research.* 2002; 16: 321-357.

[17] Batuwita R, Palade V. *Efficient resampling methods for training support vector machines with imbalanced datasets*. *International Joint Conference on Neural Networks.* Barcelona, Spanyol. 2010: 1-8.

[18] O'Fallon BD, Donahue WD, Crockett DK. A support vector machine for identification of single-nucleotide polymorphism from next-generation sequencing data. *Bioinformatics.* 2013; 29(11): 1361-1366.

[19] Sujeevan R, Hebert PD. Bold: The Barcode of Life Data System. *Mol Ecol*. 2007; 7(3): 355-364.

[20] Han J, Kamber M. Data mining: concepts and techniques. 3thed. New York (US): Morgan kaufmaann Academic Pr. 2012.

[21] Yang Yu, Liang Zhou. Acoustic Emission Signal Classification based on Support Vector Machine. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2012: 10(5): 1027-1032.

[22] Wahyuningrum, Rima Tri. Efficient Kernel-based Two Dimensional Principral Component Analysis smile Stages Recognition. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2012: 10(1): 113.

[23] Meyer D. Misc functions of the department of statistics, TU Wien. *R package version 1*. 2014: 6-3.

[24] Hsu CW, Chang CC, Lin CJ. 2003. A practical guide to support vector classification. Departemen of Computer Science and Information Engineering (TW): National Taiwan University.

[25] Yen SJ, Lee YS. Cluster-based under-sampling approaches for imbalanced data distribution. *Elsevier.* 2009; 36(3): 5718-5727.