Feature Selection Approach for Solving Imbalanced Data Problem in Single Nucleotide Polymorphism Discovery

R Nurhasanah¹, L S Hasibuan² and W A Kusuma²

¹ Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia, 20155
² Department of Computer Science, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University

*Email: rossynurhasanah@usu.ac.id

Abstract. Single Nucleotide Polymorphism (SNP) is a type of molecular marker which constitutes the phenotypic variations between individuals in certain species. In recent years, the advantages of SNP were widely considered in many fields, for instance in designing precision medicine in humans and assembling superior cultivars in plant breeding. The main challenge in SNP discovery is imbalanced data distribution between classes, where the number of true SNPs in question is much fewer than false SNPs. While the study in observing the benefit of feature selection in classification problem was widely reported, the use of this technique in solving imbalanced class problem still become interesting topic for research. In this study, we selected the features that most contribute in identifying SNP using Feature Assessment by Sliding Thresholds (FAST) method. FAST evaluates the contribution of each feature in identifying SNPs based on the Area under ROC Curve (AUC) value. SNP identification using 4 best features resulted in improved classifier performance in terms of G-Means compared to using 24 features. In addition, using feature selection techniques can reduce computational time and save resource needed.

1. Introduction
After the discovery of Next Generation Sequencing (NGS) technology, there was an explosion of genomic data which were valuable source of genetic information. NGS technology could generate a whole genome data that was gigabytes in size with a much faster time and cheaper price compared to the previous sequencing method, the Sanger method. As a comparison, the Sanger method takes 15 years in sequencing a complete human genome at a cost of more than 1 million dollars, while the NGS method only takes 2 months with cost savings of up to 100 times [1]. The rapid growth of genomic data with very large in numbers must be followed by the computational infrastructure and powerful analytical method. So that the valuable information contained in those DNA strand can be immediately extracted and utilized for human welfare. Processing these bigdata genomics is a very urgent challenge to be solved by computer scientists and data analysts. An efficient analysis method with machine learning approach is a potential solution to extract information from the genomic data [2].

One of the most important information to be explored is the existence of Single Nucleotide Polymorphism (SNP) in the DNA sequence of living things. SNP are variation of a single base in several genome sequences that control the different traits of individuals [17]. Analysis of the effects of SNP has given new knowledge for human about inheriting traits and phenotypic differences in all creatures [3]. SNP can be utilized in various fields, such as in understanding genetic factors that cause certain diseases...
in humans for example diabetes type II, down syndrome or rare diseases such as sickle cell anemia. In addition, SNP can be applied in the field of pharmacology which gave rise to a new field of research, namely pharmacogenetics. This field aims to identify SNPs related to drug metabolism. Recent research had revealed that SNP in certain genes influence the dose of certain types of drugs. Estimating the right dosage for each patient is important, and it can be done by utilizing SNP. This discovery leads to a new field in pharmacology called precision medicine or personalized medicine, which is the precise treatment for patients based on their genetic background [4].

The main problem that becomes a focus in SNP discovery is how to identify the real SNP since not all variations of nucleotides are SNP. Most of these are actually errors in the sequencing process or misalignment. This leads to an imbalanced data problem, where the positive SNP is far less than the existing errors [5]. Various research to identify SNP had been conducted before. Nielsen et al. [6] provided a general pipeline as a guide in conducting SNP discovery research. In [7], they performed data resampling using undersampling and oversampling to balance the data used. Classifiers were trained based on the Support Vector Machine algorithm. As a result, undersampling data with a ratio of 1:1 produced a better classifier and was able to identify 96% positive SNP with 84% PPV value.

Matukumalli [8] used a machine learning approach namely the C4.5 algorithm in building the SNP identifier model. Besides being good at predicting classes, C4.5 was chosen because it can produce decision alternatives or rules in the form of a decision tree based on a set of initial features. The study used 10 relevant features that have been previously validated by experts in identifying SNPs. These features are optimized by adding new features that can improve classification performance. By applying machine learning, the PPV value increases up to 10 times, reaching 84.8%.

In this study, we propose a method in SNP identification with a feature selection approach. Feature selection was known to be powerful in improving classifier performance because it only uses features that are relevant to classification and ignores irrelevant ones [9]. Feature selection is also reported to be able to overcome imbalanced data problems [10]. In this study, we use the Feature Assessment by Sliding Threshold method in selecting the features that most contribute to classification. These features are then validated by being included in the Support Vector Machine. Classification results will be analyzed with the confusion matrix to get the G-Mean value. The subset of features that produce the highest G-Mean values will be recommended to be used in classification.

2. Methods
The research method that we carried out is explained through Figure 1 below.

![Figure 1. SNP identification procedures.](image-url)
2.1. Data Collection
In this study, we used SNP candidate data from the soybean genome. Its features had been extracted by [5]. The SNP candidate data consisted of 24 features with the number of SNP candidates being 39,454,648 data. The description of the features used was explained at [7]. These data were grouped into 2 classes, namely true SNP class and false SNP class. In the feature selection process, all SNP candidate data were used to evaluate each feature. While in the sequential forward feature selection process, the data used were the SNP candidate data on chromosome number 16. Then we used a new dataset to validate the best subset features, namely chromosome number 1 from the soybean genome data. We used the selected features in building the classifier. We evaluated the ability of the classifier to identify real SNP from all SNP candidate data.

2.2. Feature Selection
Feature selection refers to a technique for extracting a features subset from the original dataset. The assumption in feature selection is that most of the information inherent in the dataset can be captured using only a small attributes subset. Reducing the number of features in a dataset can have numerous benefits such as faster model training, reduced susceptibility to overfitting and reducing storage, memory, and processing requirements during data analysis [11].

We used Feature Assessment by Sliding Thresholds (FAST) technique proposed by [10] in assessing each feature. Kumar and Bell [12] evaluated four different techniques in feature selection. The features subset produced was validated with Support Vector Machine (SVM) using various cancer datasets from NCBI. By doing statistical analysis it showed that FAST was a technique that gives better results than others, especially in classification using unbalanced data.

FAST classify each feature value on multiple thresholds and gathering statistics about the performance at each boundary. In general, the algorithm for implementing the FAST method as shown in Figure 2.

![Figure 2. FAST Algorithm](image)

Each feature values are sorted from the lowest to highest one. The sorted feature values are divided into K splits to obtain sliding thresholds. Each threshold is generated by calculating the feature values mean in every split. The number of thresholds produced is as many as the number of K that would be used as a decision boundary in doing feature analysis. While single feature classifiers usually put decision boundaries at the midpoint between two classes [10][13], the FAST technique slides the
decision boundary. So, it could increase the number of true positives found at the expense of classifying more false positives. From each decision boundary, we can calculate the true positive rate (TPR) and false-positive rate (FPR) to obtain the Receiver Operating Characteristic (ROC) value. From the ROC value, then we can calculate the area under the ROC curve (AUC) score. Since AUC is a strong predictor for performance, especially for imbalanced data classification problems, we can use this score as the feature ranking[10].

Soybean genome consists of 20 chromosomes. After implementing FAST, the 20 chromosomes obtained 20 ROC curves and 20 AUC values. SNP candidate data consists of 24 features. Each feature will be sorted according to the mean of AUC value from the highest to the lowest one. If the area under the ROC curve of a feature is wider than other features, it indicates that the feature has a greater contribution to the classification. The range of AUC scores obtained is between 0.5 and 1. If a feature is not relevant to the classification, the score will be close to 0.5, but if a feature has high relevance to the classification then the score will be close to 1. Features with higher AUC values will be selected and used in the next SNP identification process.

2.3. SNP Identification

The SNP identification was carried out using the Support Vector Machine as a classifier. The classifier is a function that maps data that is not labeled to a particular label using an algorithm. To validate the developed classifier, SNP candidate data is divided into training data and test data with k-fold cross-validation technique [14]. The $D$ dataset is divided into $k$ fold, i.e. $d_1, d_2, \ldots, d_k$ which have the same amount of data. Then the classifier is trained as many as $k$ times. For each $t \in \{1, 2, \ldots, k\}$, the classifier is trained using $(D - dt)$ data and tested using $Dt$. According to [14], 10-fold cross validation is a superior technique in choosing the best model from several model choices.

All SNP identification models in this study were built using LIBSVM e1071 in the R programming language [15] using the Radial Base Function (RBF) kernel.

2.4. Performance Metrics

The metric commonly used in measuring classifier performance is accuracy. However, with problems with unbalanced data, using accuracy can lead to inaccurate conclusions. Since the amount of negative class data is very much compared to the positive class, the classifier will get high accuracy easily. Although the classifier failed to detect the desired positive class. In this study we use the value of Geometric Mean (G-Mean) in assessing the built classifier. G-Mean is calculated based on the confusion matrix using equation 1 below.

$$G - \text{Mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}}$$

(1)

Sensitivity measures the ability of the classifier to explain data into a positive class, while specificity measures the ability of the classifier to predict negative class. Since the specificity and sensitivity values are the opposite, it means that if the specificity value rises, the sensitivity value decreases, and vice versa. So, we use G-Mean to measure classifier performance to predict negative class and positive class at once.

3. Result and Discussion

After conducting the FAST technique, we calculated the average of Area under ROC Curve (AUC) for all 20 chromosomes in soybean genome. Features with the highest AUC value are considered to have a greater contribution in predicting class. Features with AUC value > 0.5 can be seen in the following Table 1.

| Feature Number | FAST Score |
|----------------|------------|
| 3              | 0.923896518 |

Table 1. The result of FAST
To find the best feature subset that will be used in the SNP identification process, we use the Sequential Forward Feature Selection (SFFS) technique. This method searches the best subset by adding a feature into the current subset and looking at the effect of the feature on classification [16]. We compare the performance of classifiers that use 3, 4, 5 and 24 features. The results are presented in Table 2.

Table 2. Comparison of classifiers performance using 3, 4, 5 and 24 features.

| Dataset     | Sensitivity | Specificity | G - mean | Computation Time (hour) |
|-------------|-------------|-------------|----------|-------------------------|
| 3 features  | 0.59        | 0.96        | 0.75     | 2.99                    |
| 4 features  | **0.63**    | **0.96**    | **0.78** | **3.30**                |
| 5 features  | 0.61        | 0.96        | 0.77     | 7.58                    |
| 24 features | 0.60        | 0.97        | 0.76     | 28.54                   |

From Table 2 we can see that the classifier using 4 features produced the highest G-Mean value of 0.78. This means that 78% of all SNP candidate data were classified correctly by the classifier. We also observed that classifier trained using 24 features produced lower G-Mean values, only reaching 0.76. This proves that selecting features can improve classification performance because it only uses features that are relevant to classification and removes irrelevant features. Irrelevant features tend to weaken the classifier and lead to overfitting conditions. Then we can conclude that using 4 features can also shorten the time needed to do computation about 8.6 times faster than using all of 24 features available. Feature selection also reduces the dimensions of the data so that it can save the resources needed in computing. Then we can observe that from all of the classifiers built, high specificity values were obtained, which were 0.96. It means that the classifiers were able to classify 96% of negative SNP data appropriately into the negative class, and only 4% are misclassified to the positive class. This is due to the unbalanced data distribution between classes, so the classifier tends to classify samples into classes with a greater amount of data.

The four selected features are maximum quality of minor alleles (feature number 3), average quality of minor alleles (feature number 5), minor alleles frequency (feature number 7) and alleles balance (feature number 21). We use another dataset, the SNP candidate data on the first chromosome of soybean genome (Gm01) to validate these features. The performance of the classifier can be seen in Figure 3.
From the figure above, the classifier gets 96% accuracy. This means that 96% of the data can be classified correctly into their respective classes. If we observe that the sensitivity generated is 64.7%. That is, the classifier was only able to detect 64.7% positive SNPs, while 35.3% were misclassified into negative classes. The ability of the classifier to detect negative SNP was very high, indicated by a high specificity value of 98%. Meaning that of all negative SNP data, 98% could be classified correctly into negative classes. Only 2% were misclassified into positive classes. But in general, the ability of the classifier to classify samples to each class appropriately was 79.6%.

4. Conclusion
The objective of this paper is to find the most contributed features in identifying SNP of soybean genome. The results showed that the FAST method could find 4 best features that have the best AUC scores. We observed that there was slight performance improvement when implementing these selected features in classification compared to using all available features. Another benefit was since feature selection reduced the data dimension, the execution time for the classification was faster. It is not required to use all available features to get the best performances. From the experiments, we also see that there was a significant difference between specificity and sensitivity value due to imbalanced class distribution. This condition would affect the training process of SVM and tend to classify the samples to majority class.

References
[1] Kulski J K 2016 Next-Generation Sequencing — An Overview of the History, Tools, and “Omic” Applications Next Generation Sequencing - Advances, Applications and Challenges (InTech)
[2] Ma C, Zhang H H and Wang X 2014 Machine learning for Big Data analytics in plants Trends Plant Sci. 19 798–808
[3] Kusuma W A, Hasibuan L S and Istiadi M A 2016 SNPs Selection using Gravitational Search Algorithm and Exhaustive Search for Association Mapping IOP Conference Series: Earth and Environmental Science vol 31 (Institute of Physics Publishing)
[4] Bush W S and Moore J H 2012 Chapter 11: Genome-Wide Association Studies PLoS Comput. Biol. 8 1–12
[5] Istiadi M A, Kusuma W A and Tasma I M 2014 Application of decision tree classifier for single nucleotide polymorphism discovery from next-generation sequencing data Proceedings - ICACSI 2014: 2014 International Conference on Advanced Computer Science and Information Systems (Institute of Electrical and Electronics Engineers Inc.) pp 339–43
[6] Nielsen R, Paul J S, Albrechtsen A and Song Y S 2011 Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12 443–51
[7] Hasibuan L S, Kusuma W A and Suwamo W B 2014 Identification of single nucleotide polymorphism using support vector machine on imbalanced data *Proc. - ICACSIS 2014 2014 Int. Conf. Adv. Comput. Sci. Inf. Syst.* 375–9

[8] Matukumalli L K, Grefenstette J J, Hyten D L, Choi I Y, Cregan P B and Van Tassell C P 2006 Application of machine learning in SNP discovery *BMC Bioinformatics* 7

[9] Saeys Y, Inza I and Larrañaga P 2007 A review of feature selection techniques in bioinformatics *Bioinformatics* 23 2507–17

[10] Chen X W and Wasikowski M 2008 FAST: A roc-based feature selection metric for small samples and imbalanced data classification problems *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 124–32

[11] Guyon I and Elisseeff A 2006 An Introduction to Feature Extraction BT - Feature Extraction: Foundations and Applications ed I Guyon, M Nikravesh, S Gunn and L A Zadeh (Berlin, Heidelberg: Springer Berlin Heidelberg) pp 1–25

[12] Kumar P G and Bell J B B 2012 Using Continuous Feature Selection Metrics to Suppress the Class Imbalance Problem *Int. J. Sci. Eng. Res.* 3 1–9

[13] Wasikowski M and Chen X W 2010 Combating the small sample class imbalance problem using feature selection *IEEE Trans. Knowl. Data Eng.* 22 1388–400

[14] Kohavi R 1995 A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection *Int. Jt. Conf. Artif. Intell.*

[15] Meyer D, Dimitriadou E, Hornik K, Weingessel A and Leisch F 2019 e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien

[16] Setiawan D, Soroako A T, Kusuma W A and Wigena A 2018 J estr

[17] Purnami, N, Prabowo, G I, Wungu, C D K and Handajani, R 2019 Detection of single-nucleotide polymorphism Gap junction protein Beta-2 genes in deaf schoolchildren of javanese population in Surabaya, Indonesia. *Indian Journal of Otology*, 25(1), p.6.