Feature Extraction and Classification on Single Nucleotide Polymorphism

Nur Fatihah Kamarudin [a], Zuraini Ali Shah [b,*], Mohd Farhan Md Fudzee [b], Shahreen Kasim [b]

[a] School of Computing , Faculty of Engineering, Universiti Teknologi Malaysia
[b] Fakulti Sains Komputer Dan Teknologi Maklumat, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
* nurfatihah@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received: September 2019
Revised: October 2019
Accepted: November 2019

Keywords

Single Nucleotide Polymorphism, iterative prunning principal component analysis, Random Forest, Hidden Naive Bayes, Information Gain Ratio.

ABSTRACT

Malay in Peninsular Malaysia can be divided into eight sub-ethnics which are Malay Bugis, Malay, Malay Champa, Malay Jawa, Malay Kelantan, Malay Kedah, Malay Minang and Malay Pattani. Ancestry informative marker (AIM) can be used to represent the eight subethnic of Malay population in Peninsular Malaysia. In this research, single nucleotide polymorphism (SNP) datasets of eight sub-ethnics are analyses in order to obtain the AIM for Malays population in Peninsular Malaysia. However, the dataset may have outlier, missing data and redundancy that may impact the accuracy of the result. Pre-processing data is an important step that will remove the entire problem. Iterative pruning principal component analysis (ipPCA) is one of the techniques that usually use in analysis on genome datasets to extract the information. It can be applied on the high structured data and can improve the resolution of the data. It also used for structure a sub-population. Random Forest and Hidden Naïve Bayes is used to classify the SNP that can be used as AIM. Information Gain Ratio will rank the chosen AIM based on the value of each attribute

1. Introduction

Malays in one of ethnic can be found around of the world especially in Asia. One of the countries that have Malays ethnic is Malaysia. 51 percent of population in Malaysia is Malays ethnics (Cavendish and M., 2007). Malays can be divided into eight sub-ethnics which are Malay Bugis, Malay Banjau, Malay Champa , Malay Jawa, Malay Kelantan, Malay Kedah, Malay Minang and Malay Pattani. AIM can be used to be a marker that represents Malays in Peninsular Malaysia.

Ancestry informative marker (AIM) is one of the types of marker that can be extracting from single nucleotide polymorphism (SNP). SNP data is one of the big data. One SNP data may consist more that 50,000 of SNP. Analyze of the data may require large memory of computer and time consuming. So, data mining is important in analyze the data. It is used to extract the information from the big data.

Feature extraction, classification and feature selection is types of data mining. One of the algorithm can be used for feature extraction is iterative pruning principal component analysis (ipPCA). This algorithm is built based on complications of population structure analysis. There are many types of data classification for data mining. One of them is Random Forest (RF). RF has great classification performance with many ideal criteria for datasets. Hidden Naïve Bayes (HNB) is also known as one of the excellent classifier due to its accuracy and simplicity. Information Gain Ratio ranked the selected attribute based on the attribute value

2. Problem Statement

Study of single nucleotide polymorphism and ancestry informative marker is done by years but its involved other population and species. The dataset usually used in the study is HapMap, bovine, and Shriver's. There is only few study of Malays population in Peninsular Malaysia. One of the study is the genetic structure of Malay population by Hatin et al.(2011) [1]. Yet, there is no study of the ancestry informative marker of Malays in Peninsular Malaysia has been done. So, this study is aim to extract ancestry informative marker from eight ethnics of Malays single nucleotide polymorphism.

3. Objectives

The aim of this research is to identify the panel single nucleotide polymorphism, ancestry informative marker for Malays in Peninsular Malaysia. The objective of this research is

divided into three, which are:

(a) To extract the feature of the data by using iterative pruning principal component analysis

(b) To classify the ancestry informative marker by using Random Forest and Hidden Naïve Bayes

(c) To rank the ancestry informative marker by using Information Gain Ratio.

4. Methodology

In this research, the research framework is divided into four phase. Phase 1 is the phase that required an understanding on SNP and the current issue involve in SNP. This step is important in finding the highlighted issue and challenge that can lead to the purpose of this study. Additionally, the advantage of this issue has led to the analyzation of the SNP by using feature extraction and classification in order to obtain the ancestry informative marker.

The next step is carry out the resource and data collection. All the resource such as thesis, journal and lecture notes is obtained from the internet. The SNP data is obtained from Malaysian Node Human Variome Project (MyHVP) database with approval of the president[2]. The data is completely confidential and required to get permission from owner. The detail of the SNP data is explained on the next section.

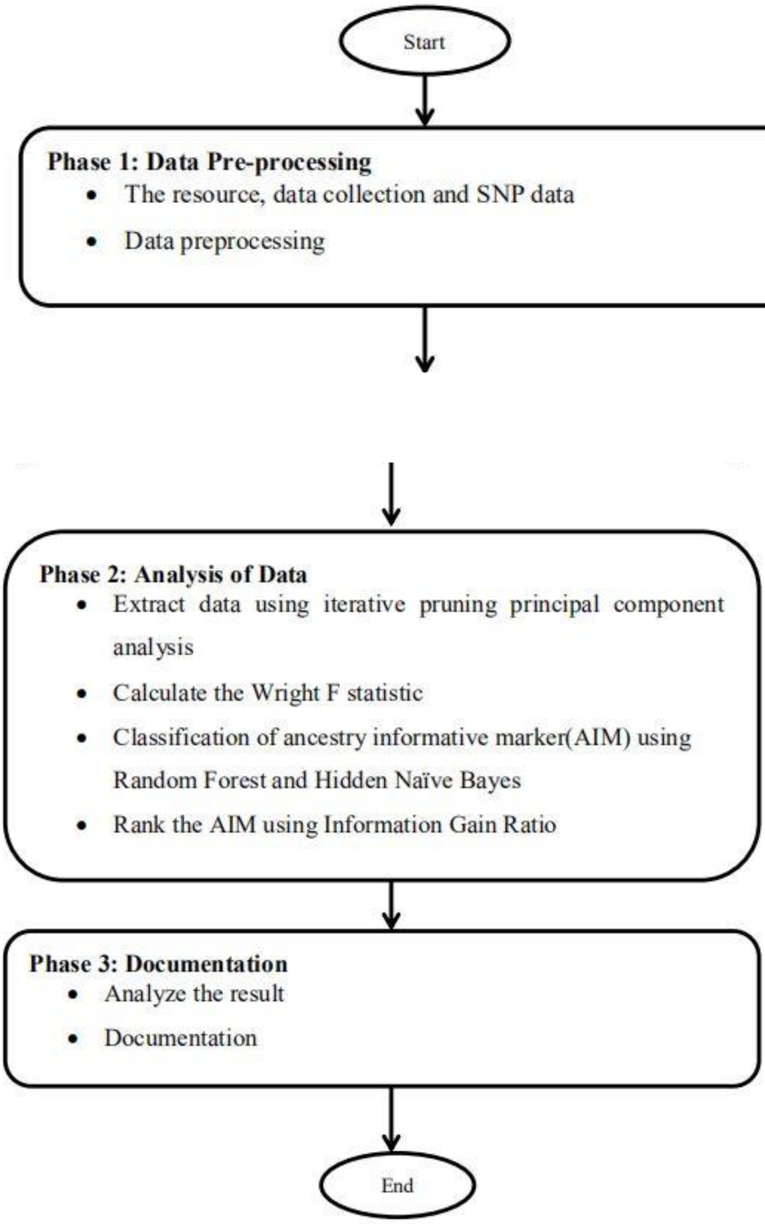
Start
Phase 1: Data Pre-processing
The resource, data collection and SNP data
Data preprocessing
Phase 2: Analysis of Data
Extract data using iterative pruning principal component analysis
Calculate the Wright F statistic
Classification of ancestry informative marker(AIM) using Random Forest and Hidden Naïve Bayes
Rank the AIM using Information Gain Ratio
Phase 3: Documentation
Analyze the result
Documentation
End

Figure 1: Research Framework

After obtain the SNP data, the data will undergo pre-processing to remove the outliers, missing data and redundancy in the data by using quality control. It also remove samples of DNA or marker that bias to the study. The pre-processing data will be process by using plink, an open source toolset for analysis of whole genome association. Phase 1 is on preprocessing. Phase 2 is analysis of the data. After the data pre-processing, the dataset will be undergo process of data analysis to obtain the information needed for the study. First analysis is feature extraction. One of the most commonly used feature extraction is Principal Component Analysis (PCA). There are many type of PCA. This study is focused one PCA methods which are iterative pruning PCA (ipPCA). IpPCA is producing scatter plot with eigen value [3]. In this step, the population is divided into sub-population by using Eigen value. The sub population is divided. The allele frequency of the data is calculated by using Wright F-statistic. It will build the genetic structure of diploid population.

The next activity on this phase is data classification. There are two classification method apply in this study which are Random Forest (RF) and Hidden Naïve Bayes (HNB). The SNP from the result obtain from feature extraction is classified into AIM. A list of AIM is obtained after the classification.

After that, the outcome of the classification is then ranked by using information gain ratio evaluation. It is rank based on its attribute value by using information gain ratio algorithm. The last phase of this study is phase of analyze the result and documentation. After that, the different between the results of each method is explained. After that, documentation is written to record all the work flow of this research. The final conclusion is based on the overall output of this research.

5. Results

Table 1: Summary of pre-processing result

Before Quality Control	After Quality Control
130 Individual ● 1 of 130 individuals removed for low genotyping ● 3 of 129 individuals removed during IBD Analysis	126 Individuals
52501 SNP ● 11357 markers to be excluded	27772 SNP

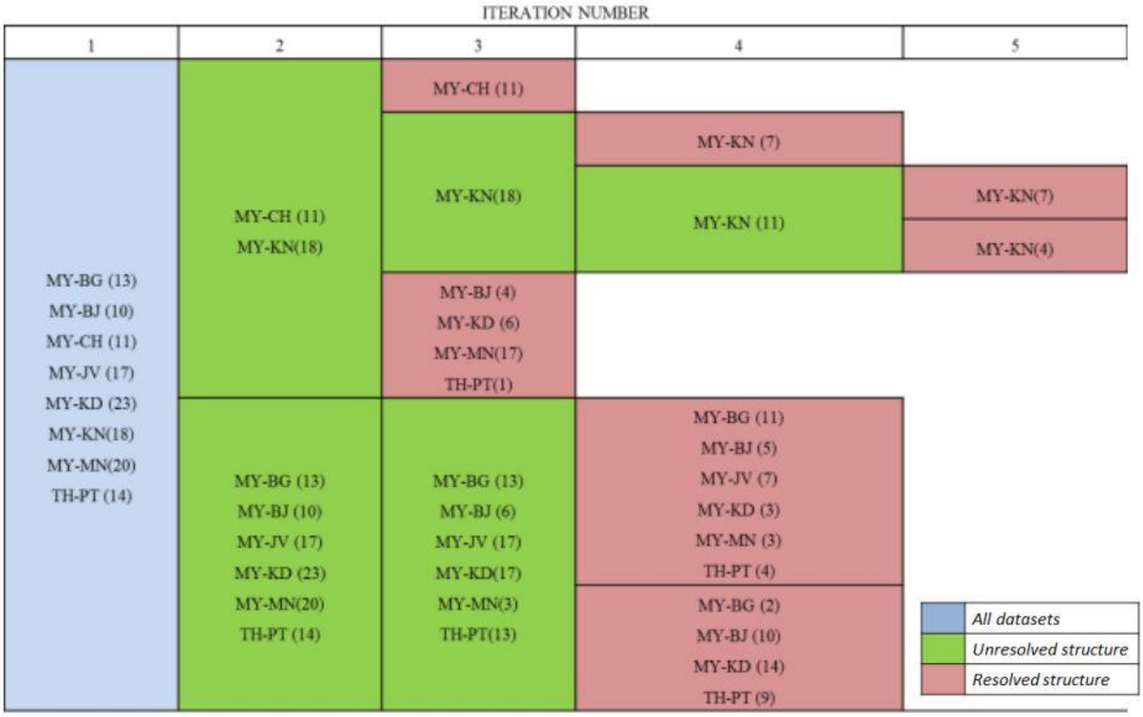

Figure 2 **: Sub-population tree obtain after run ipPCA**

Table 2: Comparison of result for Random Forest and Hidden Naive Bayes

FEATURE\CLASSIFIER	RANDOM FOREST	HIDDEN NAIVE BAYES
Mean absolute error	0.0058%	0
Root mean squared error	0.0339%	0
Relative absolute error	37.08%	0.0011%
Root relative squared error	38.2165%	0.0032%

Table 3 : top 20 of the selected SNP for AIM

No	Rs id	No	Rs id	No	Rs id	No	Rs id
1	Rs6678924_0	6	rs564367_0	11	rs1298129_G	16	rs823163_C
2	rs10492984_A	7	rs6689909_0	12	rs951241_0	17	rs1891246_A
3	rs4311853_0	8	rs10493566_0	13	rs10489823_A	18	rs10496300_A
4	rs4653014_0	9	rs1537782_0	14	rs3010367_A	19	rs3754801_G
5	rs7534892_A	10	rs10493608_0	15	rs1323012_T	20	rs10496920_A

6. Discussion

Table 1 shows the result of the pre-processing data. Before process the data using quality control, there are 52501 SNP and 130 individuals. After running the data with the QC steps, the SNP left is 27772 and 126 individuals. There are 4 individual that have been removed which one from Malay Bugis (id = MY-BG5), two from Malay Jawa (id = MYJV12 and MY-JV16) and lastly one from Malay Kedah (id = MY-KD13). The data is ready to undergo feature extraction.

Figure 2 show the sub population tree that obtained after runs ipPCA. Each cell contains ethnic's labels. The blue cell represents all dataset, green cell represent unresolved structure of nested datasets while terminated red cell shows resolved population. There are 13 nodes has been created by ipPCA process that represents as a sub-population. After this, the data were classified using RF and HNB

Table 2 shows the different between the result obtained from RF and HNB. Both of the classifier returns same number of SNP as AIM. Yet, RF accuracy is lower compared to HNB due to some error occur. Lastly, Top 20 of the SNP chosen as AIM is ranked by using Information Gain Ratio. It is rank based on the value of each attribute.

7. Conclusion

For the conclusion, pre-processing and data mining is an important steps in analyzing big data such as SNP the pre-processing used in this research has successfully deleted the missing, redundant and outliers of the data. This step is able to improve the accuracy for the next step which is feature extraction. The feature extraction used in this research is iterative pruning principal component analysis. It has constructed the structure of the Malays subpopulation. Classification is important to classify the SNP into AIM. This research prove that RF has lower accuracy compared to HNB. IGR has successfully ranked the top SNP that suitable to use as AIM.

References

[1] Hatin, W. I., Zahri, M. K., Xu, S., Jin, L., Tan, S. G., Rizman-Idid, M., ... and HUGO Pan-Asian SNP Consortium. (2011). Population genetic structure of peninsular Malaysia Malay sub-ethnic groups. PloS one, 6(4), e18312.

[2] Hashim, A. H., Etemad, A., Latif, A. Z., Merican, A. F., Baig, A. A., Annuar, A. A., ... & Shah, M. I. (2015). The first Malay database toward the ethnic-specific target molecular variation. BMC research notes, 8(1), 176.

[3] Intarapanich, A., Shaw, P. J., Assawamakin, A., Wangkumhang, P., Ngamphiw, C., Chaichoompu, K., ... & Tongsima, S. (2009). Iterative pruning PCA improves resolution of highly structured populations. BMC bioinformatics, 10(1), 1.