**Roughsets for Rice Data DNA Processing : a Data Science Approach**

Erna Budhiarti Nababan1\*, Rossy Nurhasanah1, Ade Sarah Hufaizah1

*1Universitas Sumatera Utara, Jl. Universitas No. 9, Medan, Indonesia (ernabrn@usu.ac.id)*

*Abstract –* Rice breeding programrequired to produce high quality variety of rice. DNA sequencing, annotation and mining was used to extract information about rice resistance with disease. Machine Learning approaches in bioinformatics and infrastructure are needed to process and analyzing SNP data. Roughset theory framework can be used to guarantee the validity of data attributes.

*Keywords –* DNA Sequence, Roughset.

1. **Introduction**

Oryza Sativa L or widely known as rice is a primary food with more than 50% of world population depends on it. As primary food resource, rice agriculture has been through molecular modification to produce high quality rice. *Japonica* (Nipponbare), *indica* (93-11), IR64, and *aus* is a cultigen which have been sequenced with japonica as golden standard for assembly and annotation study [1].

Genetic factors contribute significantly to determine property of high-quality and resistance to disease of the rice. An analyzer used in evaluating rice quality which can analyze starch consistency because it called Rapid Visco-Analyzer (RVA) because it only needs small number of samples to conduct data analysis, the identification of new quantitative trait loci (QTLs) for RVA was profiling of great significance to improve rice quality [5]. Research by Li [6] represent a statistical framework to Single nucleotide Polymorphisms (SNP) calling using Japonica Rice variant of Oryza Sativa to map Apparent amylose content (AAC) and Protein Component (PC) as main factor of rice disease resistance [7].

*Universitas Sumatera Utara*

**Email:** ernabrn@usu.ac.id

*Received: -----.*

*Accepted: -----.*

*Published: -----.*

© 2022. ----; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

Mahesh [1] use rice whole genome sequencing and found that *R-genes Pi-ta* and Pi54 from susceptible indica gnomes which is Tetep and Tadukan cultivar can be used as marker selection in rice breeding technology.The expansion of genomic research combining with genetics and molecular biology and advanced technological modelling used to conduct data analysis to gain another perspective.

Machine learning approaches and bioinformatics infrastructure are needed in processing and analysing SNP data. Various studies on the association of SNP with phenotypes using machine learning techniques have been conducted. In association mapping, there needs to be a framework to guarantee the validity of the variable’s sources involved in each process. Guzzetti *et al* [4] proposed a series of procedures in association of genotype data with quantitative phenotype using a regression machine learning approach. The pipeline aims to ensure reproducibility and avoid bias in selection. The conclusion obtained is that machine learning techniques can support the prediction of quantitative properties from genotype data more effectively compare with conventional methods.

Another research proposed a protocol for the steps in controlling data quality in the study of SNP associations with phenotypes [5]. The goal is to clean DNA data by removing DNA samples that can cause bias to the study. This process is a very urgent step in ensuring the quality of the data before carrying out the next process. This protocol explains the stages in quality control for each marker, including the removal of SNPs with low quality, removing SNPs that show significant deviations and eliminating SNPs that have low minor allele frequency values.

Zeng et al [6] provide an overview of the various strategies and approaches that are often used in analysing data in the genome wide association study. One of them is about data quality control, which is the most critical stage in the association mapping process. At this stage, the validity of the data will be ensured using a series of statistical procedures.

This paper consists of Introduction where we explained the background and previous research of the paper followed by our proposed method explained in section II and finally, the result obtained in this research.

1. **Methodology**

**Data Observation**.

Observasi terhadap data dilakukan sebelum melakukan data preprocessing. Dari dataset yang didapatkan pada penelitian [1], it describes unique Simple Sequence Repeat (SSR) in HR-12 (*indica*) gnome in comparison with *Nipponbare (japonica)* and93-11, dataset contains 1602 rows and 12 attributes :

* chromosome\_numbers :
* mark\_name
* SSR Motifs
* Occurrence
* Start Codon
* End Codon
* Mismatch Value
* HR12 Product Size
* Reference Product Size
* *Nipponbare* Value
* 93-11 Value
* Tetep Value
* Tadukan Value

Timeline

Description automatically generated

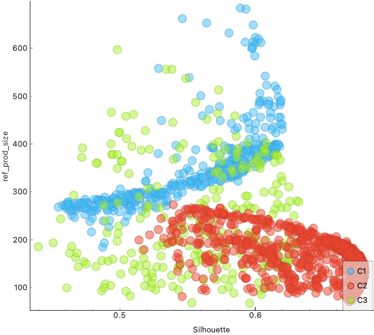
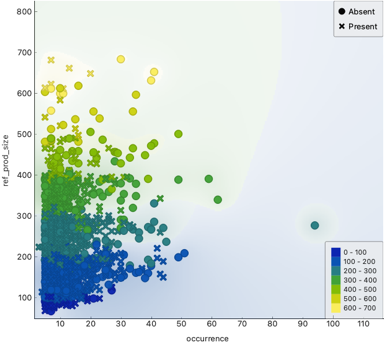
Figure 1 Histogram of Numeric Attributes

**Data Clustering**.

The data lacks target attribute used in classification task so we use clustering techniques to add one more attribute classes. Our approach is to use clustering method to cluster data based on corresponding features.

We use K-Means clustering to apply cluster class to each data as seen on table 1. Our cluster visualization can be seen on fig 2.

Figure Cluster Visualization



After we classify the cluster, now we can calculate the number of motifs in each clusters and total occurrence as seen on table 1.

Chart, histogram

Description automatically generated

Figure 3 Motif count from each clusters graph

Table 1 Number of SNP Motif Occurence on each clusters

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Count of motifs** | **Sum of occurrence** |
| C1uster 1 | 427 | 57220 |
| Cluster 2 | 940 | 102100 |
| Cluster 3 | 261 | 38460 |
| **Grand Total** | **1628** | **197780** |

**Data Transformation**.

We transform data model to show occurrence of motifs from each clusters and found 189 distinct motifs with different value from each cluster. This transformation used to find discernibility matrix and dependency weight to form a cluster.

**Rough Sets**.

Since DNA data contains collections of attributes that has correlation to each other’s we need to define which attribute has greater impact for identification. In Mathematics, sets must be exact to make reasoning possible. Pawlak proposed Rough set theory to overcome vagueness of sets mathematically with the assumption every object in data has association with others. Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The indiscernibility relation generated by characterizing similar information. The given set of functional attributes make the indiscernibility In the rough set approach is defined relative.

**Rough Sets Feature Selection**.

Imperfect knowledge commonly faced by researcher because too many irrelevant features, noise, and even misleading feature. We need to know and understand the features of the data. Feature and subset usefulness determined by relevancy and redundancy. Features that highly correlated with decision features(s) but least correlated with each other called a good feature. To discover data dependencies in a dataset, a Rough set theory (RST) can be used to reduce the number of attributes without require additional information. In perspective of dimensional reduction, predictive class attributes are those that has most informative feature [11] [12]. Research by Riza [13] has result on released packages for Rough Sets in R programming with implementation of Rough Set Theory (RST) and Fuzzy Rough Set Theory (FRST).

**Rough Sets Implementation**.

Code 1 : Implementation of Raw Datasets

##################################################

library(RoughSets)

## Read Dataset

dataset = read.csv("dataset/motifs\_cluster.csv")

## Data Prepocessing

## 1. Check and Transform Missing Value

## 2. Encoding Categorical

dataset$npb = factor(dataset$npb,

levels=c('Present','Absent'),

labels=c('1','0'))

dataset$cult\_93\_11 = factor(dataset$cult\_93\_11,

levels=c('Present','Absent'),

labels=c('1','0'))

dataset$tadukan = factor(dataset$tadukan,

levels=c('Present','Absent'),

labels=c('1','0'))

dataset$tetep = factor(dataset$tetep,

levels=c('Present','Absent'),

labels=c('1','0'))

## 3. Normalize Start and End, hr12\_pro, and ref\_prod using Z-Score

dataset$start <- (dataset$start - mean(dataset$start)) / sd(dataset$start)

dataset$end <- (dataset$end - mean(dataset$end)) / sd(dataset$end)

dataset$hr12\_prod\_size <- (dataset$hr12\_prod\_size - mean(dataset$hr12\_prod\_size)) / sd(dataset$hr12\_prod\_size)

dataset$ref\_prod\_size <- (dataset$ref\_prod\_size - mean(dataset$ref\_prod\_size)) / sd(dataset$ref\_prod\_size)

## 3. Convert dataset into Decision Table

decision.table <- SF.asDecisionTable(dataset,

decision.attr = 13,

indx.nominal = NULL)

cut.values1 <- D.discretization.RST(decision.table,

type.method = "unsupervised.quantiles",

nOfIntervals = 3)

rice.discretized1 <- SF.applyDecTable(decision.table, cut.values1)

dim(rice.discretized1)

lapply(rice.discretized1, unique)

control <- list(t.implicator = "lukasiewicz", type.relation = c("tolerance", "eq.1"),

m.owa = 3, type.aggregation = c("t.tnorm","lukasiewicz"))

summary(decision.table)

conditional.attr <- c(4,5,6,7,8,9);

IND.A <- BC.IND.relation.RST(rice.discretized1, feature.set = conditional.attr)

reduct <- FS.quickreduct.RST(rice.discretized1, control = control)

## build the decision-relation discernibility matrix

res.2 <- BC.discernibility.mat.RST(rice.discretized1, range.object = NULL)

## generate all reducts

reduct2 <- FS.all.reducts.computation(res.2)

## generate new decision table

new.decTable <- SF.applyDecTable(rice.discretized1, reduct, control = list(indx.reduct = 1))

control.1 <- list(type.relation = c("crisp"),

type.aggregation = c("crisp"),

t.implicator = "lukasiewicz", type.LU = "implicator.tnorm")

res.1 <- BC.discernibility.mat.FRST(rice.discretized1, type.discernibility = "standard.red",

control = control.1)

## generate single reduct

reduct <- FS.all.reducts.computation(res.1)

## generate new decision table

new.decTable <- SF.applyDecTable(decision.table, reduct)

Code 2 : Implementation of Clustered Motifs

#Motif Selection using FSRT

#Referensi : https://cran.r-project.org/web/packages/RoughSets/RoughSets.pdf

library(RoughSets)

dataset = read.csv("dataset/motifs\_cluster.csv", sep =";", header = TRUE)

decision.table <- SF.asDecisionTable(dataset,

decision.attr = 1, indx.nominal = 2)

reduct <- FS.feature.subset.computation(decision.table,method = "quickreduct.frst")

ds.fs <- SF.applyDecTable(decision.table, reduct)

indx <- IS.FRIS.FRST(ds.fs, control = list(threshold.tau = 0.2, alpha = 1))

selected <- SF.applyDecTable(ds.fs, indx)

…

1. **Conclusion**

Machine learning approach can be used to discover pattern in any type of data. We can discover dominant protein motifs pattern in a high quality rice DNA for further analysis. Using roughest to conduct feature selection, we may find interesting protein feature in a rice DNA specific cluster.

1. **Recommendation**

For future works, we recommend to use DNA data from wide variety of rice with some classified classes. By doing so, we may compare susceptible and resistance characteristics form of rice blast disease.

**References**

|  |  |
| --- | --- |
| [1] | H. B. Mahesh and M. D. Sirkhe, "Indica Rice Genome Assembly, Annotation and Mining of Blast Disease Resistance Genes".   1. Z. Zuo, C. Tang, Y. Xu, Y. Wang, Y. Wu, J. Qi and X. Shi Gene Position Index Mutation Detection Algorithm Based on Feedback Fast Learcning Neural Network, In Computational Intelligence and Neuroscience, 2021.   [2]  L. Clarke, S. Fairley, X. Z. Bradley, I. Streeter, E. Perry, E. Lowy, A.M. Tasse and P. Flicek, The International Genome Sample Resource (IGSR): A Worldwide Collection of Genome Variation Incorporating the 1000 Genomes Project Data , In Nucleic Acids Research, 45 D1 (2016) D854-D859.  [3]  FH. Clarke, YS. Ledyaev, RJ Stern and PR. Wolenski, Nonsmooth analysis and control theory. Springer Science and Business Media, 2008.  [4]  T.-Q. Yu, W. Jiang, T.-H. Ham, S.-H. Chu and P. Lestari, Comparison of Grain Quality Traits between Japonica Rice Cultivars from Korea and Yunan Province of China, Journal of Crop Science and Biotech, 2008 135-140.  [5]  Z. Changquan, B. Hu, K.-z. Zhu and H. Zhang, QTL Mapping for Rice RVA Properties Using High-Throughput Re-sequenced Chromosome Segment Substitution Lines, In Rice Science, 2013, 407-414.  [6]  H. Li, A statistical framework for SNP calling, mutation discovery association mapping and population genetical parameter estimation from sequencing data, In Bioinformatics, 27, 2011, 2987-2993.  [7]  Q. Liu, Y. Tao, S. Cheng, L. Zhou, J. Tian, Z. Xing, G. Liu, H. Wei and H. Zhang, Relating Amylose and Protein Contents to Eating Quality in 105 Varieties of Japonica Rice, In Cereal Chemistry, 97, 2020, 1303-1312.  [8]  H. Sakai, S. S. Lee, T. Tanaka, H. Numa and J. Kim, Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics, In Plant & cell physiology, 2013.  [9]  R. Mitchell, Web Scraping with Python: Collecting More Data from the Modern Web, California: O’Reilly, 2018.  [10]  G. Chen and X. Xie, Exon sequencing mutation detection algorithm based on PCR matching, In PLoS One, 2020.  [11]  R. Jensen and Q. Shen, Rough Set-based Feature Selection: A Review, In Rough Computing : Theories, Technologies and Applications, 2007. |