This article can be cited as H. Kuswanto, R. Mubarok and H. Ohwada, Classification Using Naive Bayes to Predict Radiation Protection in Cancer Drug Discovery: a Case of Mixture Based Grouped Data, International Journal of Artificial Intelligence, vol. 17, no. 1, pp. 186-203, 2019.

Classification Using Naive Bayes to Predict Radiation Protection in Cancer Drug Discovery: a Case of Mixture Based Grouped Data

Heri Kuswanto[1], **Rizky Mubarok**[2] **and Hayato Ohwada**[3]

[1]Department of Statistics, Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember (ITS)
Kampus ITS Sukolilo, Surabaya 60111, Indonesia
heri_k@statistika.its.ac.id

[2]Department of Statistics, Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember (ITS)
Kampus ITS Sukolilo, Surabaya 60111, Indonesia
mubarokrizky06@gmail.com

[3]Department of Industrial Administration, Faculty of Science and Technology
Tokyo Universty of Science, Chiba-Japan
ohwada@rs.tus.ac.jp

ABSTRACT

One treatment for cancer that is widely used is radiation therapy or radiotherapy using compounds that kill cancer cells.The effectiveness of the radio theraphy is assessed from the percentage of cancer cell death rate. This research examines 84 compounds where each compound is composed by 217 features leading to high dimensionality of the data. Feature selection is carried out based on the mean value of Gini (MDG) and it is able to sort the most important features used in the classification using Naive Bayes. The Naive Bayes has a weak performance to classify the raw dataset i.e. using threshold of 10% cancer cell death rate. A grouping based on mixture distribution found 30% cell death rate as a new threshold, and it improves the performance of Naive Bayes both in training and testing dataset evaluated using AUC (Area Under Curve). The optimal classification for testing dataset is obtained by using either 20% or 25% most important features with AUC close to 60%, where it is about 15% higher than classification using threshold of 10%. Meanwhile, the AUC of training dataset reached more than 70%.

Keywords: Mixture, Naive Bayes, Cancer, Radiotheraphy.

2010 Mathematics Subject Classification: 62P10, 68T05, 68T10.
2012 Computing Classification System: Computing methodologies - Machine learning - Learning paradigms - Supervised learning - Supervised learning by classification .

1 Introduction

Cancer is a general term for a large group of diseases characterized by the growth of abnormal cells which can then attack parts of body around these abnormal cells and/or spread to other organs. Cancer can affect almost all parts of the body and have many molecular and anatomical subtypes, each of which requires a specific treatment strategy. Cancer cells can be found in malignant tumors that have the characteristics of cells growing indefinitely, have a sheath, and spread or can infiltrate the surrounding tissue. Handling cancer will be better if it can be detected early. Hence, prevention efforts are required by increasing public awareness in recognizing the symptoms and risks of cancer in order to determine the appropriate preventive measures. Various studies related to cancer treatment have been conducted by researchers in the world, including (Diamantis and Banerji, 2016) who examine Antibody-Drug Conjugates (ACDs) as a new class that arises from cancer treatment that combines selectivity from targeted treatment with cytotoxic potency from chemotherapy drugs. (Sharma, Plummer and Stock, 2016) examined drug combinations and radiotherapy to improve the clinical outcomes of patients with cancer.

In the past few years, machine learning has become an interesting spotlight in the field of health and drug discovery, one of which is cancer. Various studies related to the use of machine learning in the field of health and drug discovery, especially in cases of cancer have been carried out by researchers in the world, including the research conducted by (Soria, Garibaldi, Biganzoli and Ellis, 2008) who compared the C4.5 method, multilayer perceptron classifier (MLP), and Naive Bayes in breast cancer cases where MLP produced the highest accuracy of 94.9% compared to Naive Bayes and C4.5 with accuracy of 93.1% and 87.6%. Comparison of the C4.5 and Naive Bayes methods was also conducted to classify survivability of heart cancer patients by (Dimitoglou, Adams and Jim, 2012) where C4.5 with an accuracy of 94.44% is better than Naive Bayes with an accuracy of 92.38%. (Kathija and Nisha, 2016) also conducted a breast cancer data classification study using SVM and Naive Bayes, and revealed that Naive Bayes outperforms SVM with accuracy of 97%. In addition, the Naive Bayes method, logistic regression, and decision tree were also compared by (Mandal, 2017) in the case of breast cancer cell detection where logistic regression outperforms the others. A study was also conducted by (Elsayad and Elsalamony, 2013) who compared the performance of several decision tree models (C & R, CHAID, QUEST, C5.0) and SVM to diagnose breast cancer, the results showed that SVM had the highest accuracy among others. (Aruna, Rajagopalan and Nandakishore, 2011) also conducted a study to detect breast cancer using the Naive Bayes method, RBF neural networks, J48, CART, and SVM-RBF Kernel. In this research, the SVM-RBF produces highest accuracy than the others. More recent study using ensemble based classification methods were carried out by (Kuswanto and Werdhana, 2018). All of those researches revealed that the performance of the method depends on the case, which means that there is no single method which always outperforms the others.

Radiation therapy or commonly known as radiotherapy is one of the treatments for cancer. This therapy is still used even though it has adverse side effects, namely normal cell death and nor-

mal tissue around P53-induced cancer cells (Morita, Ariyasu, Wang, Asamaru, Onoda, Sawa, Tanaka, Takahashi, Togami, Nenoi, Inaba and Aoki, 2014). P53 is a tumor suppressor gene that acts to stop tumor development. This is done by activating several proteins that trigger the death of damaged cells so that the cells do not replicate uncontrollably. It is the background of study carried out by (Ariyasu, Sawa, Hanaya, Hoshi, Wang and Aoki, 2014) who conducted the study by developing 84 compounds which were then tested on normal cells and cells affected by gamma radiation. The first experiment was carried out by administering compounds to normal cells which was able to measure the toxicity of these compounds. While the second experiment was carried out by giving compounds to cells that had been exposed to gamma radiation (10 Gy) to measure radiation protection. Cell death rates were used as indicators in both trials. If the cell death rate in the first trial is low and in the second experiment is high, then the compound can be used as a radioprotector. Data obtained from these studies were then used for further research conducted by (Matsumoto, Aoki and Ohwada, 2016) who compared several machine learning methods i.e. random forest (RF), support vector machines (SVM), extreme gradient boosting (XGB), and K-nearest neighbor (KNN) to optimize radiation protection and toxicity. The study revealed that all of these methods has relatively low accuracy to predict radiation protection compounds with about 60% accuracy rate. The machine learning approach was applied to the most important features selected by the Gini Index.Dealing with the issue of low classification accuracy as shown in (Matsumoto et al., 2016), one of the potential reasons is the threshold used to group the class response. The study used 10% cell death rate as the threshold to characterize a low and high radiation protection, determined by treating the radioprotection as a bivariate case with the level of toxicity. This threshold does not provide a significant and clear boundary between the group leading to a low classification accuracy. Consequently, the compounds with cell death rate around 10% will be easily misclassified.

Our preliminary study found that the data has an indication of cluster, which is not necessarily to be exactly two clusters. If there are indications that some groups emerge from the data, the distribution can be a mixture (Dempster, Laird and Rubin, 1977). This research proposes a new idea of classifying the radio protection where the new threshold is identified by a data driven approach i.e. mixture distribution approach. The mixture distribution will found the optimum number of cluster, and hence the definition of low and high radiation protection will be changed following the newly defined threshold, which is statistically justified. The mixture distribution provides a justification on the significant different of the class response group, and hence it eliminates the confusion of membership category for any observation induced by unclear boundary. Moreover, (Jung Kim, 2015) found that mixture distribution for threshold classification is an optimal procedure for a classification involving a nonlinear classification rule due to its complex distribution.

The classification of the new class of the response category is carried out by Naive Bayes. The Naive Bayes classifier is used because this method only requires a small amount of training data to determine the estimated parameters needed in the classification process (Pattekari and Parveen, 2012). Moreover, the performance of Naive Bayes has not been investigated in

(Matsumoto et al., 2016). Nevertheless, several researches showed that Naive Bayes is one of the good classification methods in term of the accuracy. In this paper, simple optimization rule is used to optimize the Naive Bayes Classifier. Meanwhile,the parameters of Gaussian Mixture Model is optimized by using Expectation-Maximisation (EM) algorithm (see (McLachlan and Krishnan, 2008); (Dempster et al., 1977)). In fact, there have been many optimization algorithms developed in the literature. (Pozna, Precup, Tar, Skrjanc and Preitl, 2010) proposed heuristic modelling algorithm expressed in terms of homogenous combinations of the classical system dynamics and the Bayesian degree of truth employed in modelling, which increase the transparency and alleviated the computational time. (Vascax, 2012) proposed Fuzzy cognitive maps to overcome the limitation on conventional rulebased systems on describing of complex dynamic systems requiring nontrivial decision procedures. (Saadat, Moallem and Koofigar, 2017) did a comprehensive simulation study on the performance of Echo States Network (ESN) combined with different optimization algorithms and found that ESN combined with Heuristic algorithm outperform the other methods such as Recursive Least Square and Particle Sworm Optimization. (Vrkalovic, Lunca and Borlea, 2018) used Gey Wolf Optimizer Algorithm on the Takagi-Sugeno fuzzy controllers are designed.

In this paper, the compounds to optimize radiation protection will be classified based on cancer cell death rates which has the characteristic of high dimensional data. Basically, the Naive Bayes is not designed for high dimensional data, and hence, feature selection is carried out prior to applying the classification method. The features are selected by Mean Decreasing Gini (MDG) to determine the importance of each feature (feature importance), as part of Random Fores (RF) approach. This index has been proven by (Calle and Urrea, 2010) to be a more accurate and stable method for calculating the importance feature. Several different percentages of the most importance features will also be examined. Based on the results of these classifications, the compounds that have high radiation protection can be recommended for use as a good radioprotector.

The structure of the paper is as follows. The next section provides a brief description about the supporting theories including the feature selection using Mean decreasing Gini, Naive Bayes classifier and Gaussian Mixture Model. The optimization procedure as well as the algorithm are described. Section 3 describes the research methodology e.g. the data and steps of the analysis. In section 4, we perform the results of the analysis. It begin with the discussion of the feature selection and is continued with the classification using Naive Bayes applied to raw dataset. Furthermore, we discuss the results of grouping the response class using mixture distribution as well as Naive Bayes performance applied to the newly grouped data. Section 5 concludes the paper.

2 Supporting Theories and Algorithms

This section briefly describes the theory and algorithm of Mean Decreasing Gini used for feature selection, the Naive Bayes Algorithm as well as the Gaussian Mixture Model and its esti-

mation procedure.

2.1 Feature Selection Using Mean Decreasing Gini

Feature selection is a way to reduce the dimension of the variable (or feature), especially in the case of high dimensional data. One of the ways to select features is by calculating the feature importance. Mean Decreasing Gini (MDG) is a measure used to calculate the feature importance level resulted from a Random Forest method. The detail about Random Forest algorithm is omitted in this paper. Readers interested to learn more about random forest are referred to (Breiman, 2001) and (Shi and Horvath, 2006), or (Liaw and Wiener, 2002) for the implementation of Random Forest in R software. (Calle and Urrea, 2010) found that the MDG is better than Mean Decreasing Accuracy (MDA) and Gini Index to measure the importance measure because MDG is more stable and yield on more robust result. The Mean Decrease Gini (MDG) is the average total decrease of node impurity for a certain feature, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest. The MDG measures of how important a feature is for estimating the value of the target feature across all of the trees that make up the forest. The more importance the feature, the higher the MDG. The formula (2.1) below is used to calculate the MDG:

$$MDG_p = \frac{1}{s}\sum_{t=1}^{s}[d(p,t)I(p,t)] \tag{2.1}$$

where $d(p,t)$ is the decreasing level on Gini Index of feature x_p on t-th node, $I(p,t)$ has value 1 if it splits the t-th node and 0 if the others, s is the number of node in the tree, and x_p is the p-th feature.

In Random Forest algorithm depends on two tuning parameters i.e. the number of trees (**ntree**) and the number of preselected directions for splitting (**mtry**). The MDG for each feature is calculated using those two parameters which yield on the highest accuracy. (Scornet, 2018) provides a comprehensive study about tuning parameters in random forest. In summary, the algorithm of random forest can be described as follow:

Algorithm 1: Random Forest

Input: A training set $D_n := (x_1, y_1), ..., (x_n, y_n)$, features ***F***, and number of trees (**ntree**).
  function RANDOMFOREST (D_n, F)
    S $\leftarrow \emptyset$
    for $i \in 1, ...,$ **ntree do**
      $D^i \leftarrow$ **mtry** bootstrap sample from D_n
      $s_i \leftarrow$ RANDOMIZEDTREELEARN(D^i, F)
      $S \leftarrow S \cup s_i$
    end for
    return S
  end function

function RANDOMIZEDTREELEARN(**C,F**)
At each node:
$f \leftarrow$ very small subset of **F**
Split on best feature in f
return The learned tree
end function

2.2 Naive Bayes Classifier

Naive Bayes classifier is one of the most popular classification methods and listed as among best ten algorithms in data mining according to (Wu, 1983). Given a set of variables, $\mathbf{X} = \{x_1, x_2, ...x_p\}$, the Naive Bayes classification is based on simple probability on Bayes Theorem assuming independence among variables (features) as follow:

$$P(C|x_1, x_2, ..., x_p) = \frac{p(C)p(x_1, x_2, ..., x_p|C)}{p(x_1, x_2, ..., x_p)} \tag{2.2}$$

where C represents the set of categorical level, $p(C_j|x_1, x_2, ..., x_p)$ is the posterior probability of class membership i.e. the probability that **X** belongs to C_j. The algorithm of Naive Bayes can be described as follows:

1. Let the training sets of samples is denoted by T and there are j classes for each label C such that $C = \{C_1, C_2, ..., C_j\}$. $\mathbf{X} = \{x_1, x_2, ..., x_p\}$ represents an p dimensional vector of sample for the corresponding p measured values of features, $A_1, A_2, ..., A_p$ respectively.

2. Given the sample $\mathbf{X}$, the highest conditional posterior probability is used as a classifier to predict the class. $\mathbf{X}$ is predicted to belong to the class C_j if and only if

   $$P(C_j|\mathbf{X}) > P(C_i|\mathbf{X}) \qquad for \quad 1 \leq j \leq p, i \neq j$$

   the goal is to find the class that maximizes $P(C_j|\mathbf{X})$. By using the Bayes theorem in ((2.2)), the class C_j for which $P(C_j|\mathbf{X})$ is maximized is called the maximum posteriori hypothesis.

3. The maximization is applied to $P(\mathbf{X}|C_j)P(C_j)$ as $P(\mathbf{X})$ is the same for all classes. If $P(C_j)$ is unknown, then it is assumed that the classes are equal such that $P(C_1) = P(C_2) = ... = P(C_j)$, and therefore $P(\mathbf{X}|C_j)$ is maximized. Meanwhile, if $P(C_j)$ is known, then $P(\mathbf{X}|C_j)P(C_j)$ is maximized. The $P(C_j) = freq(C_j, T)/|T|$ can be used to estimate the class of prior probability.

4. Computing $P(\mathbf{X}|C_j)$ will be very computationally intensive in case of high dimension of the data sets (i.e. containing of many features). To deal with this, the Naive Bayes applies naive assumption of the class conditional independence during the computation of $P(\mathbf{X}|C_j)P(C_j)$. This assumption leads to the values of the features that are conditionally independent of one another, given the class label of the sample such that

   $$P(\mathbf{X}|C_j) \approx \Pi_{k=1}^{p} P(x_k|C_j)$$

The probabilities $P(x_1|C_j), P(x_2|C_j), ..., P(x_p|C_j)$ can easily be estimated from the training set. Recall that here x_k refers to the value of feature A_k for sample $\mathbf{X}$, and the following rules are applied:

- In case that A_k is categorical, then $P(x_k|C_j)$ is the number of samples of class C_j in T (denoted as $freq(C_j, T)$) having the value x_k for feature A_k, divided by $freq(C_j, T)$.
- In case that A_k is numerical, it is assumed that the values have a Gaussian distribution as follow

$$P(x_k|C_j) = g(x_k, \mu_{C_j}, \sigma_{C_j})$$

The μ_{C_j} and σ_{C_j} are the mean and standard deviation of values of attribute A_k for training sample of class C_j.

5. Two parameters that need to be optimized are $P(\mathbf{X}|C_j)$ and $P(C_j)$. Therefore, for predicting the class label of $\mathbf{X}$, $P(\mathbf{X}|C_j)P(C_j)$ is evaluated for each class C_j. The class label of $\mathbf{X}$ is predicted to be C_j if and only if the class is the one that maximizes $P(\mathbf{X}|C_j)P(C_j)$ such that

$$C \leftarrow \arg\max_{C_j} P(C_j)\Pi_k P(\mathbf{X}|C_j) \tag{2.3}$$

The algorithm above is the most simple method assuming that the class probabilities and conditional probabilities are known variable. Therefore, simple rule by choosing the maximum posterior probability can be applied. In case those two parameters are treated as unknown variable, several optimization techniques can be applied such as Combination of the Gradient and Newton (CGN) methods. More detail about CGN can be found in (Taheri and Mammadov, 2013).

2.3 Gaussian Mixture Model

The Gaussian Mixture Model (GMM) can be seen as a soft K-means clustering (Hastie, Tibshirani and Friedman, 2008). The GMM with two mixtures is given as an illustration. The Gaussian Mixture Model (GMM) is written as

$$\begin{aligned} p_{Y_i}(y_i) &= \sum_{c_i} (\pi_c N(y_i; \mu_c, \sigma^2))^{1(c_i=c)} \\ &= qN(y_i; \mu_1, \sigma^2) + (1-q)N(y_i; \mu_2, \sigma^2) \end{aligned} \tag{2.4}$$

where q is the mixing proportion or mixture weight satisfying $\sum_c q_c = 1$. Each Gaussian density has a mean μ_c and standard deviation σ. The log-likelihood of the parameters derived from its joint density of all observations is

$$\ln p_{Y_1^n}(y_1^n) = \sum_{i=1}^{n} ln(\pi_1 N(y_i; \mu_1, \sigma^2) + \pi_2 N(y_i; \mu_2, \sigma^2)) \tag{2.5}$$

In this case, we have four parameters to be optimized i.e. $\theta = \{\pi, \mu_1, \mu_2, \sigma\}$. The parameter estimation is done by maximizing the log-likelihood $\log p_Y(y, \theta)$. After some mathematical manipulations such as marginalisation over *C* as well as using Jensens's equality, we obtain

$$\log p_Y(y, \theta) \geq E_{qc}\left[log\frac{p_{Y,C}(y, C, \theta)}{q_C(C)}\right] \tag{2.6}$$

The M-step in EM algorithm maximizes ((2.6)) with respect toθ yields on:

$$\hat{\theta} \leftarrow \arg\max_{\theta} E_{qc}[logp_{Y,C}(y, C; \theta)] \tag{2.7}$$

Meanwhile, the E-step maximizes (xx) with respect to q_c so that

$$\hat{Q}_C(.) \leftarrow p_{C|Y}(.|y; \theta) \tag{2.8}$$

In summary, the algorithm can be written as follows:

Algorithm 2: Expectation-Maximization for Gaussian Mixture Model

---

Inputs: Given Observation y with $p_{Y,C}(y, c; \theta)$ and $p_{C|Y}(c|y; \theta)$ as the joint distribution and conditional distribution respectively, and initial values $\theta^{(0)}$

function EM($p_{Y,C}(y, c; \theta)$, $p_{C|Y}(c|y; \theta)$, $\theta^{(0)}$)

for iteration $t \in 1, 2, ...$**do**

$q_C^t \leftarrow p_{C|Y}(c|y; \theta^{(t-1)})$ (**E-step**)

$\theta^(t) \leftarrow \arg\max_\theta E_{q_C^t}[p_Y, C(y, C; \theta)]$ (**M-step**)

if $\theta^(t) \approx \theta^{(t-1)}$ **then**

return $\theta^(t)$.

The algorithm above stops when the value of the Log-likelihood reached its convergence. With respect to the number of mixture, the optimum mixture distribution is the one with highest Log-likelihood value. The guarantee that the EM algorithm will monotonically converge to local optima has been proven by (Dempster et al., 1977) and (Wu, 1983). (Naim and Gildea, 2012) showed that the setting of initial values $\theta^{(0)}$ only influences the convergence speed.

3 Research Methodology

3.1 Data

The data analyzed in this paper is secondary data produced by (Ariyasu et al., 2014), and hereafter we called the data as raw dataset. The data contains some compounds that are related to the suppressor gene P53. In these data, there are 84 cancer drug compounds compiled by 217 predictors (hereafter denoted as features). The experiment was carried out by giving compounds to cells that had been exposed to gamma radiation (10 Gy). this experiment was able to measure radiation protection. If the cancer cell death rate in the experiment is high, then the compound can be used as a radioprotector. Table 1 below listed the response class as well as sample of features names.

Table 1: Variable

Variable	Variable Name
Y	Class target
	Y(0) = Low protection radiation (cancer cell death rate 10%)
	Y(1) = Low protection radiation (cancer cell death rate >10%)
X1	pKa
X2	ALogP 98
X3	ALogP MR
.	.
.	.
.	.
X217	Zagreb

3.2 Steps of Analysis

The analysis on this research is conducted by the following steps:

- Carry out feature selection by calculating the feature importance using Mean Decreasing Gini (MDG) with the following percentages : 5%, 10%, 15%, 20%, 25%, 30%, 35% and 100% of the total number of features.
- Perform classification of compounds based on two classes of radiation protection (raw dataset) using Naive Bayes. In this step, the data is divided into 10 parts (k=10), and then calculate the probability for every j-th feature, where j = 1,2,..., p. The probability is calculated using normal distribution. Based on this, the posterior probablity for every category of the compound is calculated to label the class. Furthermore, calculate the classification accuracy, sensitivity and AUC.
- Group the class response (cancer cell death rate) based on mixture identification. The number of mixture will be set as c=2,3 and 4. Furthermore, calculate the log-likelihood for each number of mixture. The posterior probability will also be calculated to assign the data into the new class. The same steps as described in the previous step are applied to each K.
- Compare the performance of the Naive Bayes to classify the cancer death rate using raw class and mixture based class.

All analysis above are carried out by using packages in R language programming. The dataset as well as the program codes to generate outputs in this analysis is available from the author upon request.

4 Results and Discussions

4.1 Feature Selection

This research applies Mean Decreased Gini (MDG) to select the feature due to high dimensionality of the dataset. The MDG measures the feature importance based on Gini Impurity Index used in splits calculation within the training set. The important features will be selected based on the 5%, 10%, 15%, 20%, 25%, 30%, 35% and 100% of the total features. The 5% to 35% represent the case where the data is no more in high dimensionality, while 100% is examined to see the performance of the method in a high dimensional case. The MDG is derived from the random forest approach. It uses two parameters i.e. **mtry** (number of variable that is randomly drawn as candidate in every split) and **ntree** (number of tree), where the values are predetermined by considering the computational burden. A research by (Oshiro, Perez and Baranauskas, 2012) studied about the number of tree and concluded that larger number of trees does not always improve the performance of the forest. Moreover, (Scornet, 2018) showed that **mtry** size and **ntree** play similar roles in random forest performance, where both increase linearly with the computational time. In this study, the candidate of parameters used to calculate MDG from random forest are 1, 2 to 20 (for m**try**) and 100, 200 to 1000 (for **ntree**). The optimum **mtry** and **ntree** is the one with highest accuracy. Figure 1 below depicts the accuracy obtained from the combination of both tuning parameters.

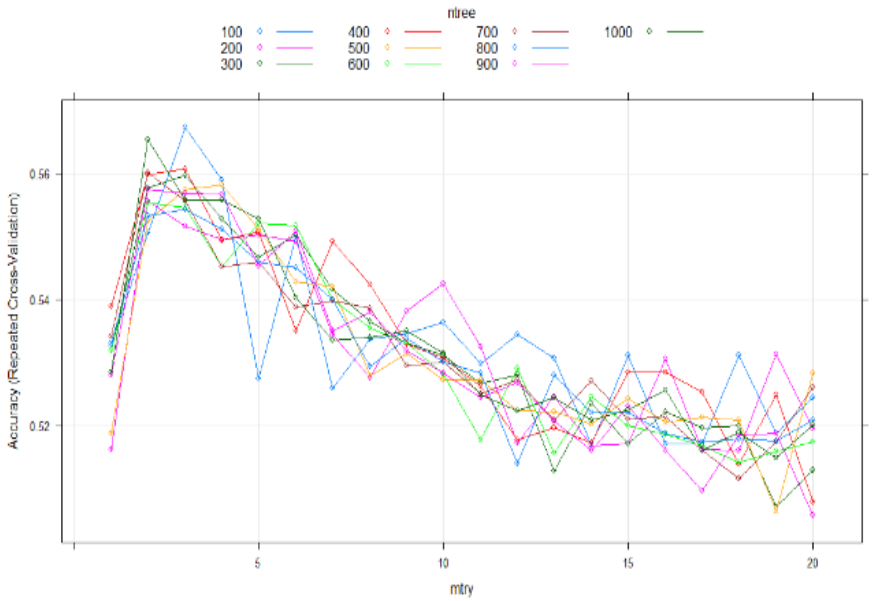


Figure 1: Tuning parameters

Figure 1 reveals that the optimal accuracy is about 58% reached under **mtry** 3 and **ntree** 100. Using these parameters, the MDG can be calculated to see the importance of each variable. The accuracy tends to decrease by increasing the **mtry**. The results of MDG calculation for the 20 most important features are listed in Table 2. We see that the 91th feature (Jurs_FNSA_1)

Table 2: Mean Decreasing Gini of 20 most important features

Variable	Feature Name	MDG
X91	Jur_FNSA_1	0.582
X48	Molecular_SASA	0.573
X120	Minimized_Energy	0.509
X6	ES_sum_aaN	0.487
X87	Dipole_Z	0.479
X141	SAscore	0.473
X134	Shadow_YZ	0.460
X45	Molecular_Fractional Polar Surface Area	0.459
X73	Kappa_1_AM	0.455
X40	QED_Unweighted	0.449
X70	JX	0.436
X43	Sascore_Fragments	0.434
X11	Jurs_TASA	0.426
X119	Energy	0.422
X107	Jurs_RPCS	0.415
X5	ES_sum_aaCH	0.414
X114	Jurs_WNSA_3	0.408
X57	CHI_V_0	0.396
X69	IC	0.393
X67	AIC_Mean	0.393

is the most important feature, and 48th feature (Molecular_SASA) is ranked as the 2nd most important feature. Meanwhile, 207th and 209th features are the least important features (the output is omitted here). In this case, both features will only be involved in the classification with 100% features. Classification with 5% top features will use only 11 most importance features.

4.2 Classification Using Naive Bayes on the Raw Dataset

The first step of the classification is to examine the training dataset. Furthermore, the model in training set will be used to classify the testing data into two classes i.e. low and high radiation protection. Both training and testing dataset are divided into 10-fold cross validation, meaning that the total accuracy, sensitivity and specificity as well as AUC of every fold are averaged. Classification with Naive Bayes produces the probability of a compound in every class, and then the highest probability is selected between both classes to classify the compound. The performance of Naive Bayes classifier can be seen in Table 3.

From the table, we see that the performance of Naive Bayes on training dataset is higher than testing dataset. In fact, the proportion of class response is relatively balance i.e. 0.46 for class 1 and 0.54 for class 0. In this case, the total accuracy can be used to assess the classification performance. The AUC is performed to make it comparable with the mixture based group in

Table 3: Performance of Naive Bayes classifier applied to raw dataset

% feature	Data	Total Accuracy	AUC
5%	Training	0.613	0.594
	Testing	0.572	0.549
10%	Training	0.610	0.559
	Testing	0.553	0.537
15%	Training	0.632	0.623
	Testing	0.489	0.477
20%	Training	0.619	0.610
	Testing	0.514	0.499
25%	Training	0.617	0.606
	Testing	0.525	0.512
30%	Training	0.618	0.607
	Testing	0.536	0.522
35%	Training	0.639	0.630
	Testing	0.560	0.551
100%	Training	0.598	0.579
	Testing	0.489	0.472

case the class proportion is imbalance. Increasing the percentage of the features does not improve the performance. Nevertheless, the Naive Bayes is robust against number of feature shown by the slight changes of the total accuracy and AUC among the number of feature settings. In summary, the performance of Naive Bayes classifier applied to testing dataset can be seen in Figure 2.

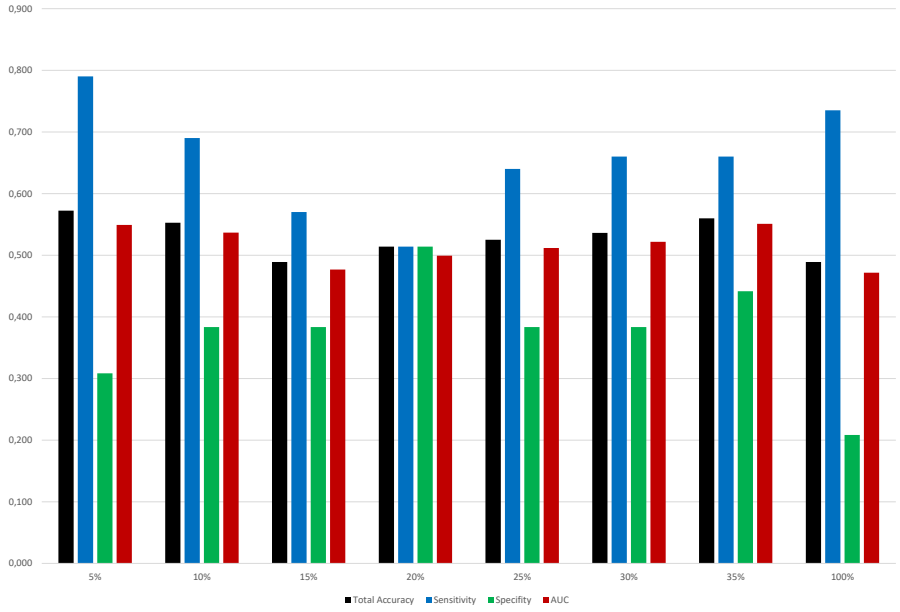
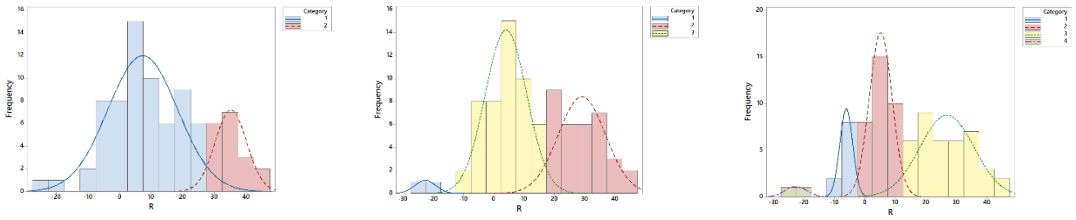


Figure 2: Performance of Naive Bayes Classifier on testing dataset

Based on the chart in Figure 2, we see that using 5% most importance features will lead to the highest total accuracy and AUC i.e. 0.572 and 0.549 respectively. These values indicate a low predictability as only 57.2% of the compounds can be correctly classified by Naive Bayes classifier. However, the sensitivity values is high indicating that the Naive Bayes has good ability to correctly classify the compound into class 1. The next subsection examines the performance of Naive Bayes applied to the class response grouped by mixture distribution. It is expected that the there will be an improvement in those performance criterias.

4.3 Grouping the Response Class Using Mixture Distribution

The classification using Naive Bayes applied to the raw dataset resulted on relatively low accuracy, in line with the results of classification conducted by (Matsumoto et al., 2016). This part takes a look on the data carefully and carries out the data driven exploration using normal mixture distribution. Figure 3 below presents the distribution of the data estimated with three different number of mixtures (2,3, and 4).



(a) (b) (c)

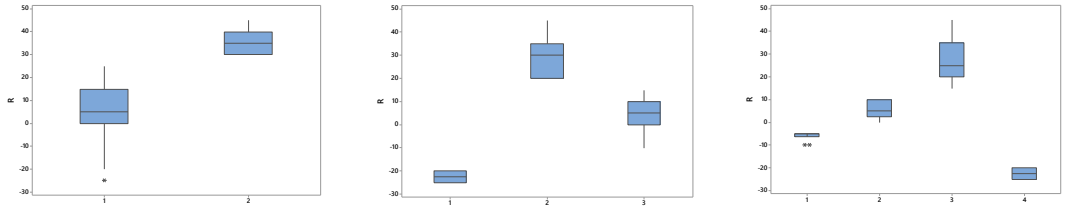
Figure 3: Distribution Plot (a) K = 2 (b) K = 3 (c) K = 4.

Table 4: Performance of Naive Bayes classifier applied to raw dataset

K	Group 1	Group 2	Group 3	Group 4
2	66	18		
3	2	33	49	
4	10	45	27	2

From the figure, we notice that using threshold 10% does not provide clear separation of the group. The log-likelihood values for those three different number of mixtures are -344.26. -341.81 and -340.44 respectively. Although 4 mixtures has the highest log-likelihood, these values are not significantly different and hence, the analysis will be preformed for all mixtures. The number of feature in each class is summarized in Table 4.

We see that there is a serious imbalance proportion of the class response for all thee mixtures. This research does not apply any specific treatment dealing with this imbalance issue. However. the assessment on the performance can be based on AUC which takes into account the imbalance issue. The summary of the category can be seen from the boxplots in Figure 4.



(a) (b) (c)

Figure 4: Boxplot of cancer cell death rate (a) K = 2 (b) K = 3 (c) K = 4.

The boxplots showed that using K=2, the group (class) one has average cell death rate 7.42% obtained from the compound with cell death rate under 30%, while group two has average cell death rate 35.28% representing the compound with cell death rate above 30%. With K=3, the first group represents a condition where the cell death rate between -25% to -20%, group two has cell death rate between -10% to 15%, while group three for cell death rate above 20%. The 4 mixtures grouped the cell death rate of -25% to -20% as group one and so forth. We see from the figure that the groups formed by mixture distribution have no overlap observation. On mixture with 3 and 4 groups, there are compounds with only negative cell death rate. It means that using the features during radiotheraphy will harm the patients as they tends to grow the new cancer cell. The compounds with negative cell death rate are listed in Table 5.

4.4 Classification Using Naive Bayes on Mixture Based Grouped Data

Table 6 shows the performance of Naive Bayes applied to the newly defined class of cancer cell death rate. Due to the imbalancing issue of the response category, the assessment will be

Table 5: Compounds with negative cell death rate

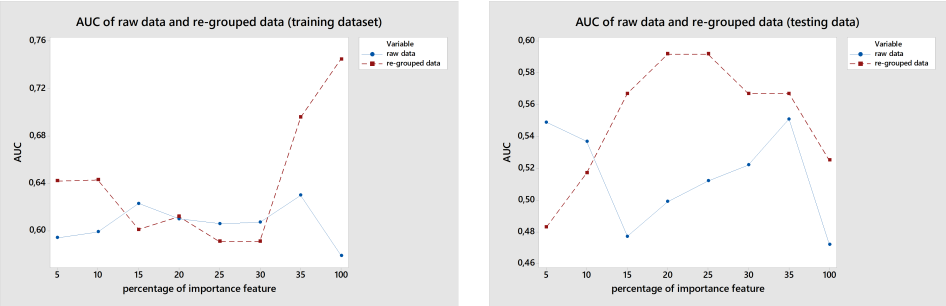
Variable	Compounds
-25%	ST-1
-20%	AS-3
-10%	KH-23, MH-9
-5%	AS-4, KT-1, MH-14, SAr-5, UM-8, YM-14, YN-4, YN-6

focused on AUC values as listed in Table 6.

Table 6: Performance of Naive Bayes classifier applied to raw dataset

% feature	Data	2 Classes	3 Classes	4 Classes
5%	Training	0.642	0.536	0.381
	Testing	0.483	0.356	0.439
10%	Training	0.643	0.458	0.381
	Testing	0.517	0.356	0.462
15%	Training	0.601	0.429	0.378
	Testing	0.567	0.367	0.458
20%	Training	0.612	0.487	0.369
	Testing	0.592	0.300	0.439
25%	Training	0.591	0.458	0.394
	Testing	0.592	0.333	0.371
30%	Training	0.591	0.444	0.386
	Testing	0.567	0.333	0.371
35%	Training	0.696	0.444	0.386
	Testing	0.567	0.356	0.394
100%	Training	0.745	0.437	0.403
	Testing	0.525	0.333	0.508

Based on the values in the table, we see that setting the class into three and four would not improve the classification performance as the AUC for all cases are under 50% and far below the AUC performed in Table 3. Meanwhile, the results of classification using two classes improves the classification performance. If we compare the AUC in the last column of Table 3 (corresponding to two class responses with threshold of 10% cell death rate) with the values in the third column of Table 6 (corresponding to the threshold of 30% cell death rate), we observe that there is an improvement in some cases as shown in Figure 5 below.



(a) (b)

Figure 5: AUC of raw data and re-grouped data (a) training dataset (b) testing dataset

The figures show that in most cases, the classification using the re-grouped data increases the accuracy both in training and testing set. For the training dataset, there is a significant improvement when the classification is done using all features, where the AUC reaches about 75%. Meanwhile, for the testing dataset, a significant improvement happen when the classification is done using 20% or 25% of the important features, reaching about 60% accuracy.

5 Conclusion

This research proposed a new group of class response for the cancer cell death rate obtained from the mixture distribution, where the new classes are defined as below and above 30% cell death. The classification using Naive bayes indicated that two classes yield on optimal result, while three and four classes have very low predictability. There is a significant improvement both in training and testing phase after regrouping the class response into two classes with threshold of 30%, assessed with AUC. This research showed that the class response threshold is an important factor that may influence the performance of the classicifation. In order to optimizing the predictability of the cancer death rate, this research suggested to use 20% or 30% of the important features. The Jurs_NFSA and Molecular_SASA are two most important features to predict the cancer cell death rate. The Naive Bayes is relatively robust against the number of features.

Acknowledgment

This work was supported by the Funding from the Ministry of Research and Higher Education Indonesia through the Research Grant for International Collaboration and Scientific Publication Scheme. The authors wish to acknowledge the anonymous reviewers for their detailed and constructive comments to the manuscript.

References

Ariyasu, S., Sawa, A., Hanaya, K., Hoshi, M., Wang, B. and Aoki, S. 2014. Design and synthesis of 8-hydroxyquinoline-based radioprotective agents, *Bioorganic and Medicinal Chem-*

istry **22(15)**: 3891–3905.

Aruna, S., Rajagopalan, S. P. and Nandakishore, L. P. 2011. Knowledge based analysis of various statistical tools in detecting breast cancer, *Computer Science and Information Technology* **2**: 37–45.

Breiman, L. 2001. Random forests, *Machine Learning* **45(1)**: 5–32.

Calle, M. L. and Urrea, V. 2010. Stability of random forest importance measures, *Briefings in Bioinformatics* **12(1)**: 86–89.

Dempster, A. P., Laird, N. M. and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm (with discussion), *Journal of The Royal Statistical Society B* **39**: 1–38.

Diamantis, M. and Banerji, U. 2016. Antibody-drug conjugatesan emerging class of cancer treatment, *British Journal of Cancer* **114(4)**: 361–367.

Dimitoglou, G., Adams, J. A. and Jim, C. M. 2012. Comparison of the c4.5 and naive bayes classifier for the prediction of lung cancer survivability, *Journal of Computing* **4(8)**.

Elsayad, A. M. and Elsalamony, H. A. 2013. Diagnosis of breast cancer using decision tree models and svm, *Journal of Computing* **85(3)**: 19–29.

Hastie, T., Tibshirani, R. and Friedman, J. 2008. *The elements of statistical learning*, Vol. 2 of *Springer Series in Statitsics*, Springer Verlag, Berlin heidelberg.

Jung Kim, H. 2015. A best linear threshold classification with scale mixture of skew normal populations, *Computational Statistics* **30(1)**: 1–28.

Kathija and Nisha, S. 2016. Breast cancer data classification using svm and naive bayes techniques, *International Journal of Innovative Research in Computer and Communication Engineering* **4(12)**: 21167–21175.

Kuswanto, H. and Werdhana, R. 2018. Classification of alzheimer related genes using lorens with important and significant features, *Internetworking Indonesia Journal* **10(1)**: 29–34.

Liaw, E. and Wiener, M. 2002. Classification and regression by randomforest, *R News* **2(3)**: 18–22.

Mandal, S. K. 2017. Performance analysis of data mining algorithms for breast cancer cell detection using naive bayes, logistic regression, and decision tree, *International Journal of Engineering and Computer Science* **6(2)**: 20388–20391.

Matsumoto, A., Aoki, S. and Ohwada, H. 2016. Comparison of random forest and svm for raw data in drug discovery: prediction of radiation protection and toxicity case study, *International Journal of Machine Learning and Computing* **6(2)**: 145–148.

McLachlan, G. and Krishnan, T. 2008. *The EM algorithm and extensions*, Vol. 2 of *Wiley Series in Probability and Statistics*, Wiley, New York.

Morita, A., Ariyasu, S., Wang, B., Asamaru, T., Onoda, T., Sawa, A., Tanaka, K., Takahashi, I., Togami, S., Nenoi, M., Inaba, T. and Aoki, S. 2014. As-2, a novel inhibitor of p53-dependent apoptosis, prevents apoptotic mitochondrial dysfunction in a transcription-independent manner and protects mice from a lethal dose of ionizing radiation, *Biochemichal and Biophysical Research Communications* **450(4))**: 1498–1504.

Naim, I. and Gildea, D. 2012. "convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients", *"Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK"* pp. 1–8.

Oshiro, T., Perez, P. and Baranauskas, J. 2012. *How many trees in a random forest?*, Vol. 7376 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin Heidelberg.

Pattekari, S. A. and Parveen, A. 2012. Prediction system for heart disease using naive bayes, *International Journal of Advanced Computer and Mathematical Sciences* **3(3)**: 290–294.

Pozna, C., Precup, R. E., Tar, J. K., Skrjanc, I. and Preitl, S. 2010. New results in modelling derived from bayesian filtering, *Knowledge-Based Systems* **23(2)**: 182–194.

Saadat, J., Moallem, P. and Koofigar, H. 2017. Training echo state neural network using harmony search algorithm, *International Journal of Artificial Intellegence* **15(1)**: 163–179.

Scornet, E. 2018. Tuning parameters in random forests, *ESAIM: Proceedings and Surveys* **60**: 144–162.

Sharma, R. A., Plummer, R. and Stock, J. K. e. a. 2016. Clinical development of new drugradiotherapy combinations, *Nature Reviews Clinical Oncology* **13**: 627–642.

Shi, T. and Horvath, S. 2006. Unsupervised learning with random forest predictors, *Journal of Computational and Graphical Statistics* **15(1)**: 208–222.

Soria, D., Garibaldi, J. M., Biganzoli, E. and Ellis, I. O. 2008. "a comparison of three different methods for classification of breast cancer data", *"Seventh International Conference on Machine Learning and Applications, San Diego"* pp. 619–624.

Taheri, S. and Mammadov, M. 2013. Learning the naive bayes classifier with optiization models, *International Journal of Applied Mathematics and Computer Science* **23(4)**: 787–795.

Vascax, J. 2012. Adaptation of fuzzy cognitive maps by migration algorithms, *Kybernetes* **41(3-4)**: 429–443.

Vrkalovic, S., Lunca, E.-C. and Borlea, I.-D. 2018. A comparison of three different methods for classification of breast cancer data, *International Journal of Artificial Intellegence* **16(2)**: 208–222.

Wu, C. j. J. 1983. On the convergence properties of the em algorithm, *The Annals of Statistics* **11(1)**: 95–103.