Received: November 7, 2019. Revised: December 12, 2019.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Optimization of Feature Selection Using Genetic Algorithm in Naïve Bayes Classification for Incomplete Data

Bain Khusnul Khotimah[1,2] Miswanto Miswanto[3]* Herry Suprajitno[3]

[1]*Faculty of Science and Technology, University of Airlangga Surabaya, Indonesia*
[2]*Departmentof Informatic Engineering, University of Trunojoyo Madura, Indonesia*
[3]*Department of Mathematics, University of Airlangga Surabaya, Indonesia*
* Corresponding author's Email: miswanto@fst.unair.ac.id

Abstract: In the case of high dimensional data with missing values, the process of collecting data from various sources may be miss accidentally, which affected the quality of learning outcomes. a large number of machine learning methods can be applied to explore the search area for imputation and selection of features and parameters. ML classification needs preprocessing with self-organizing map imputation (SOMI) before the imputation of missing values is done to improve the accuracy of the model. This study introduces a new approach that combines naïve Bayes classification (NBC) and genetic algorithm (GA) optimization procedures to effectively explore the search space based on a sample of experimental points. GA is a classification model approach based on the selection of features that cause computational problems, such as reduced dimensions, uncertainty and imbalanced data sets with various classes. In the experiment, preprocessing the data using SOMI yielded error results that were up to 10% for various data sets with missing data compared to other methods. In the SOMI-GANB hybrid model, the experimental results show that the proposed method can significantly improve accuracy by up to 90% compared to other imputation methods and without feature selection. SOMI can be used for homogeneous, heterogeneous and mixed data sets. The results from the experiment clearly showed that the proposed method could significantly increase the yield compared to the other imputation methods and without feature selection. The combination of GA and naïve Bayes classification was chosen because they are simple, easy-to-understand methods that are very effective in finding optimal solutions from a set of possible solutions. Naïve Bayes imputation had higher accuracy compared to neural network imputation.

Keywords: Incomplete data, Feature selection, SOMI, Genetic algorithm, Naïve bayes.

1. Introduction

Data sets generally have missing values and other deficiencies. Irrelevant features can be identified to reduce computational complexity [1-4]. Machine learning (ML) requires preprocessing to handle characteristic deficiencies such as missing and inconsistent values. Sets of collected data can have several other deficiencies, such as values that are non-discredited, incomplete, noisy, etc. Missing values in a set are considered problematic if they have a large influence on the decision making. Some of the ML methods that can be applied to overcome missing values in large heterogeneous data sets are Neural Network Imputation, SOMI (self-organizing map imputation), k-NNI (k nearest network imputation), SVMI (support vector machine imputation), NBI (naïve bayes imputation), ensemble classifiers, decision tree, etc. [5]. Other computational methods for dealing with the problem of missing values use the mechanism of evolution to find substitute values, such as genetic algorithm (GA), ant colony optimization (ACO), particle swarm optimization (PSO), artificial classification to optimize the parameters in a variety of applications [4, 6, 7]. Genetic algorithms can find a diverse set of solutions with search techniques based on the evolutionary principles of natural selection and genetics because of their ability to search various regions in the solution space. Computing in an imputation process gives several estimates for missing values and simulates a large number of withdrawals from the

population to estimate unknown parameters. The imputation of missing data has several advantages, such as producing estimates that are not biased towards missing values, keeping the natural variability of the observed data, and providing a measure of the uncertainty introduced by the missing data [8, 9].

ML in classification containing missing values is based on a priori knowledge or statistical information extracted from patterns in search space depending on the use of variables (feature vectors). Classification requires data preprocessing, such as selecting the best feature combination using appropriate groupings on the training set and the testing set [10, 11]. Clustering is used in imputation methods such as Imputation K-Means, FCMI (fuzzy c-means imputation), and SOMI to find replacement values based on the weight of the cluster [12]. SOMI methods are based on the concept of object distance and the weight of the training results. SOMI is trained using a data set without missing values and then provides imputed data for a second data set with missing values as prediction samples [13,14]. SOMI can provide reasonable estimates without statistically significant differences based on expert judgment, so it is suitable for imputation and can save considerable computation time. SOMI for multiprocessing on high-performance clusters reduces the need for adjustment of learning levels, momentum values, kernels and extensive activation functions to speed up overall processing. The best attributes need to be selected for imputation to reduce computational complexity [15].

Naïve Bayes classification (NBC) can be used to provide values for mixed data sets with missing values in two categories of variables, namely discrete and continuous variables [16]. NBC has weaknesses such as the probability function not being able to measure the accuracy of its predictions, because the probability value in NB depends on feature diversity, weight optimization and feature selection for classification [17]. In addition, it has problems with missing data that often occur in the training and testing data sets so that the resulting error is greater. The missing values must be filled in by another method. Common methods for completing preliminary data in NBC are the mean and the mode methods [18, 19]. In this study, process combined clustering, classification and optimization techniques. SOMI is a clustering technique that fills in missing data using a weight generated during the learning stage. Self-organizing mapping (SOM) is a popular artificial neural network method that can be applied for various purposes, including clustering and classification for high-dimensional data visualization [20]. The advantage of using SOMI is that it can be used to group continuous data as well as categories. Thus, SOMI is suitable to complement NBC's preprocessing in classifying mixed data. However, NB has very sensitive weaknesses in the selection of features, so it weighting and independent variable selection are required to improve model accuracy [21]. The genetic algorithm is an iterative method to get a global optimum for the selection of the features to be used as input for the naïve Bayes process [22-28]. This research used SOMI combined with feature selection by an evolutionary algorithm and NBC. Evolutionary algorithms were specifically developed to handle problems with high-dimensional data to reduce processing time so that the outcome is more quickly obtained. The most commonly used evolutionary algorithms are genetic algorithms because they can reduce the number of attributes in high-dimensional data without reducing the information from the data. Features selection must be done before classification using the appropriate heuristic information and classifiers to obtain an optimal feature set using training and testing data sets. A method for selecting significant features is proposed here that combines a genetic algorithm and the Bayesian theorem to estimate missing values. This study developed a genetic algorithm to optimize feature selection by classifying high-dimensional data with high computational efficiency and a very high accuracy rate. Genetic algorithms with classic optimization techniques perform poorly, especially when many features are missing.

This paper is organized as follows. Section 1 explains the background of the issue of missing values. Section 2 presents the imputation process with SOMI. In Section 3, we briefly introduce GA Feature Selection, a hybrid technique combining naïve Bayes and GA. In Section 4, we present the data preprocessing method. Finally, Section 5 presents the results of the experiment and our conclusions and recommendations for future research

2. SOM imputation

A self-organizing map provides a mapping of the input data from a higher dimensional space, *d*, to a lower dimensional space, *ld*. Basic SOM consists of nodes placed in a lower dimensional array with the weight of each node in the higher dimensional weight vector representing the input data. Nodes that are spatially near the array have the same weight vector. For each training input vector *x*, the neuron with the weight vector that has the greatest similarity with *x* is called the best matching unit (BMU) [13-15]. The

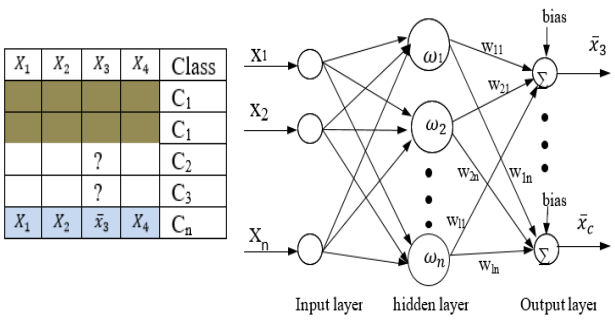

Figure 1. Imputation of missing data with SOMI

environment function applies SOM imputation to handle missing values using the features with missing values as input for certain maps and ignoring some missing variables when the distance between observations and nodes is calculated. The SOM based imputation model is illustrated in Fig. 5 for each input vector with missing values, the node is used to measure the distance with known attributes. Each missing value is replaced with the weight vector of an attribute [20].

Imputation of missing value$x_{i_m}(t)$ is replaced by the weight at the feature position and the same time sequence as the data position. If the missing data on feature 3 and class C = 2, then the imputation process can be seen in the following equation:

$$. x_{i_m}(t) = \bar{\omega}_{ij}(t) \tag{1}$$

$$\bar{x}_3 = \bar{\omega}_{3j} \tag{2}$$

In Fig. 1 there are four attributes, where attribute x_3 is incomplete. For example, the use of SOM with imputation of the x_3 attribute will replace the missing values with estimated values in the output of the weight of the same class of data. $X_{tot} = (X_1, X_2, \dots, X_n), (X_{m1}, X_{m2}, \dots, X_{mn}), \dots, (X_{n1}, X_{n2}, \dots, X_{nt})$ are multivariate variables *n*; X_{mn} are individual elements missing random variables. So, X_{nt} represents the second database containing all data; and *t* is the effective time sequence. The imputation mechanism uses SOM to estimate missing variable values using prototypes in the form of weights in a group of appropriate SOM learning outcomes. SOMI is a direct substitution process using the final weights of the components of the most similar prototype vector (BMU) and substituting with an average weight value corresponding to the SOM BMU prototype vectors of a number of its neighboring units [29].

3. Evolutionary algorithm for feature selection problems

Evolutionary algorithms are important for solving combinatorial problems. They include genetic algorithms (GA), ant colony optimization (ACO), particle swarm optimization (PSO), artificial bee colony optimization (ABC), bat algorithm (BA), fire bat algorithm etc. [24, 25]. In [30], a merged model was built to improve the ability to optimize features and parameters in electronics and communication, data mining applications. Genetic algorithms are a natural approach to selecting feature subsets represented in the form of bit-strings. The settings for each bit indicate whether or not the feature is appropriate [31]. Feature selection can experience difficulties due to interactions between subsets of variables that are mutually interdependent and redundant. The weakness of GA is that it may prematurely find solutions that are less than optimal for highly fit individuals who dominate the population at an early stage. To counteract this, feature weighting methods have been introduced [30]. The simplest approach for evaluating feature subsets is using complexity by selecting features, including evaluating features with Decision-Tree or Naïve Bayes classification [31]. The use of very few features or unrepresentative features can lead to misclassification, while having too many features can cause 'dimensional curse' problems, because it increases the cost of data acquisition and computational complexity [31]. A complete search procedure that will determine the best subset of *N* available features requires checking all subsets. This approach uses a heuristic search using particular classifier error probability criteria [32]. *N* features that represent patterns in search space as vectors $\bar{x} = (x_1, x_2, \dots, X_n)$. Feature selection finds the best part of features *y* to improve the performance of optimal criterion function *J* (.) using classification accuracy $\bar{y} = (y_1, y_2, \dots, y_t)$. The applied feature selection method finds the best *y* features in order to optimize classification performance by applying certain criteria functions $J(.)$. The result is a new feature vector with a lower dimension for $M()$. The mapping function is such that $\bar{y} = M(\bar{x})$ specifies optimal performance by fitness function $J(.)$. The result of applying *M* is making *y* such that $|y| \leq |x|$ and increasing class separation in the feature space is defined by x [32].

$$\bar{y} = M(\bar{x}) \tag{3}$$

$$J(\bar{y}) = \min(\forall \bar{y})\, J(\bar{y}); \\ J(\bar{x}) = \min(\forall M(\bar{x}))J\{M(\bar{x})\} \tag{4}$$

Among these feature selection methods, GA is commonly used to get solutions to various combinatorial problems. GA can be combined with various classifications for optimization of features and parameters, such as NB, NN, DT, SVM. In this paper, a combination of GA and NB is proposed to handle classification with heterogeneous missing values in large datasets [26-28].

4. Genetic algorithm

Genetic algorithms perform the optimum value search process at several points simultaneously in one single generation [1]. The iteration process is carried out with an evolutionary generation-to-generation approach, but the number of chromosome members with the best fitness for each generation will be maintained because it is a set of solutions [24]. Chromosomes can be binary, integer or decimal codes. In the process of evolution, a number of genes that make up the chromosome will undergo a process of crossover and mutation [25]. Genetic algorithms use probabilistic transition to select the best chromosomes, which are kept alive by the best-fitness function to obtain the optimum solution [30, 31]. The process of evaluating the fitness of each individual chromosome is done by changing the genotype of the chromosome to its phenotype. Binary strings are converted to variables in the form of pairs of real numbers [*m, n*]. The initial population is a binary string of length *n*, called the individuals. The selection process maintains it until the next generation with the roulette wheel method. Thus, a chromosome with a high fitness value has a greater chance of being selected. Selection of the individuals that will move on to the next generation is done by generating random numbers $r \in (0,1)$ [29]. The process stages of a genetic algorithm are shown in Fig. 2.

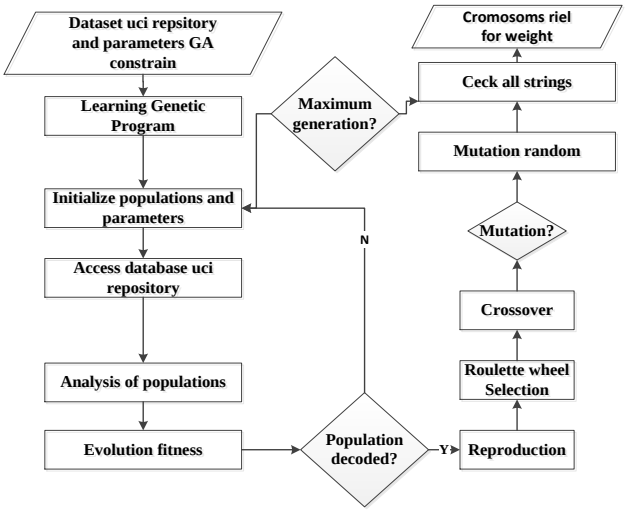

Figure 2. The process of new generation on Genetic Algorithms

Crossover uses the one-cut point method by randomly pairing chromosomes, then selecting 1 crossover point to determine the chopping position in the chromosomes. For example, the 2nd and 5th chromosomes are chosen as the first pair to cross, and $r = 2$ is the crossover point. In this research, choosing $\alpha = 0.01$ means that around 1% of genes in the population will mutate. If the value of gene *i* is 0 then the value of the gene will change to 1, whereas if the value is 1 then it will change to 0. Furthermore, the genes on the 1, 3 and 4 chromosomes with a value of 1 mutate by changing into genes with a value of 0. Chromosomes with a high fitness value will have a high probability of reproducing in the next generation [30].

4.1 Fitness function

The GA approach calculates the fitness of the population of chromosomes (strings) that represent a combination of features from the solution set, which requires an evaluation function, namely fitness function *F (.)*. The algorithm manipulates a set of chromosomes in a particular population, with operator mechanisms such as crossover, inversion, and mutation. Data with mixed attributes produce a fitness function that is calculated based on the selected features from *i* features among the total features. The length can vary according to the size of the total number of features *n* and the number of features selected. The chromosome length is the same for each chromosome. The fitness function evaluates the chromosome population that represents a combination of features to produce a series of solutions. The features in the fitness function use the mixture of attributes that represents the average classification accuracy. Fitness functions can choose individuals by tournaments and operators to produce elite chromosomes to produce new populations [28, 31]. The fitness function uses the cross validation accuracy of the classifier and is trained to select the feature subset that represents the accuracy of the data (x).

$$fitness(x) = accuracy(x) \tag{5}$$

$$accuracy = \frac{TN+TP}{TN+TP+FN+FP} \tag{6}$$

Based on the value of True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) can be obtained the value of accuracy, precision

and recall. Accuracy values describe how accurately the system can classify data correctly.

4.2 Naïve bayes genetic algorithm hybrid approach

Naive Bayes (NB) is a simple probabilistic classification that sums the combination of frequencies and values in an existing dataset. The algorithm uses the Bayes theorem to calculate the probability of all independent attributes given by the value of the class variable [33]. The advantage of using naive Bayes is that it only requires a small amount of training data to determine the estimated parameters needed in the classification process. Naive Bayes works very well in most complex real-world situations in handling missing values in homogeneous or heterogeneous data sets [34]. The Bayes formula is carried out by computing $(C|\, X_1, X_2, \dots, X_n)$ using multiplication rules. If there is a missing value for one of the variables $x_1 = x_{mis}$ then variable *C* is replaced with the value with the posterior probability calculated by NBC:

$$P(X_1|Y,C) = \frac{(Y,C|X_1).P(F_1)}{P(Y,C)} \quad (7)$$

$$P(X_1|X_2,X_3,X_4) = \frac{P(X_2,X_3,X_4|X_1).P(X_1)}{P(X_2,X_3,X_4)} \quad (8)$$

Preprocessing using SOMI is necessary in hybrid NB and genetic algorithm models to simplify learning [13]. The function of the spatial pattern feature to proceed to space classification $P \to G \to C$ is $F = X\ x\ W \to Y \to K$, where X is the data space pattern; W is the set of feature weights; Y is the feature space; and K is the set of class labels. Feature selection is done by determining weight vector *w** and then transporting the pattern features from space to space to improve the performance of the NB classifier. $x = [x_1, x_2, \dots, x_i, \dots, x_N]$, and $x' = [x_1, x_2, \dots, x_i', \dots, x_N]$, with $x_i \neq x_i'$ by replacing missing values with weights *w* later $F(x,w) = F(x',w)$. Thus, the posterior probability NBC in the complete data is:

$$P(C|X_1, X_2, \dots\ X_n) = \frac{P(C)P(X_1, X_2, \dots, X_n|C)}{P(X_1,X_2,\dots,X_n)} \quad (9)$$

with Variables C representing classes and variables in $X_1, X_2, \dots\ X_n$ represents a feature that explains the characteristics of the data needed, new NBC:

$$P(C_k|X) = \frac{P(C_k)\prod_{i=1}^{n} P(X_i|C_k)}{P(X)} \quad (10)$$

$$P(C_k|X) = \frac{P(C_k)\prod_{i=1}^{n} P(X_1, w_2, \dots, X_n|C)}{P(X_1,w_2,\dots,X_n)} \quad (11)$$

The probability of NBC is obtained by entering certain characteristic samples into class *C* as posteriors, i.e. priors multiplied by the probability of the characteristics of class *C* as the probability of the characteristics of the sample occurring globally [12, 13]. NB with feature optimization using GA is a hybrid NB model and feature selection method with removal of redundant and irrelevant features. The selection process finds the relevant features and leaves out irrelevant features, which is expected to improve the reliability of the classifier. The task of the first genetic algorithm is to encode the chromosomes in the *n* features used for classification. The sample election feature that is the weakest is removed from the subset of size *k* from the previous step, namely (f1, f5, f6, f7), by iteratively evaluating the subset of smaller (f1, f5, f7), (f1, f5, f6), (f5, f6, f7) and (f1, f6, f7). Here, we assume that the best performance subset of size 3 is (f5, f6, f7). Chromosome *n* is selected as the initial population of the GA by using the crossover and mutation operations. Then, the obtained feature is encoded in binary strings with a length of *n* bits. Bit '0' means that the feature is not selected, while bit '1' means that the feature is selected. Examples of chromosomes can be seen in Table 1.

Illustration of chromosome groups representing features with *F* = index of features, *n* = total number of features, an *ns* = number of selected features.

Table 1. Feature selection process

Features	F_1	F_2	F_3	…	F_{20}	…	F_n
Chromosome (*ns*)	1	0	1	…	0	…	**1**

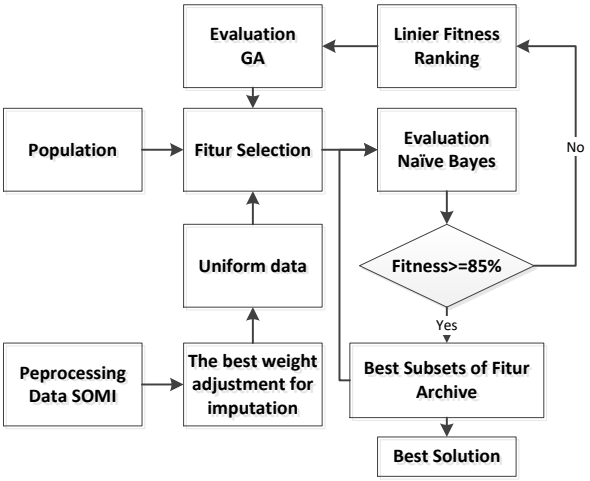

Figure 3. General process of hybrid preprocessing, feature subset selection and classification using SOMI/GA/NB

The proposed model is shown in fig. 3 as the hybrid SOMI procedure begins with preprocessing to handle data variation, to make the data set suitable for mixed data classification. This approach GA was related for feature selection for the UCI machine learning dataset missing value general, which resulted in accuracy above 85%. These were used as a reference in the calculation of fitness for classification [25-27].

5. Results and discussion

The training instance is represented by x_t, which indicates the value for the set of attributes, time and class. The attribute is represented by $\{a_1, a_2, \dots, a_n\}$ and the class is denoted by $\{C_1, C_2, \dots, C_n\}$ for the specified value. The missing values are added to the set of instances of attributes with missing values in the original data set. The domain values of the attributes as the populations are used as the set of solutions for choosing attributes after imputation of the missing values. In this method, the initial stage conducts preprocessing with SOMI, followed by the hybrid GA and NB model approach to carry out feature selection. The results of the first scenario are shown in Table 2 for small-size data sets with the number of features $f < 19$, medium-size data sets with $20 < f < 49$, and large-size data sets with $f > 50$. In the first scenario, *n* features were selected using the genetic algorithm, while 448 features were obtained from the feature extraction process [34**].** In addition, the dimensions covered by the data set have a large spectrum, ranging from 10 to 100. Glass, Heart Disease and Record Linkage Comparison Patterns were small-size data sets. WDBC was a medium-size data set was and the last two data sets, Sonar and Numeral, were large in size.

Table 3 shows several imputation algorithms applied to both high and low-dimensional data sets. Mean imputation finds the average of certain variables with complete data to replace each row of features that correspond to their average position. Modus imputation is used for a set of discrete data and shows better results than mean imputation, especially for sparse data [35]. Furthermore, mean and median imputation is used to find the middle value of all known data values to replace the other values. Hot deck is an improvement of missing data imputation using average, modus or median mode to provide better results compared to deleting incomplete data or when compared to imputation methods with constant values [3, 38]. SOMI outperformed the mean, modus and median Hot Deck methods and outlier detection is also done efficiently. The SOMI algorithm can have low means square error in the weight of each element of the replacement features while maintaining a high convergence speed [13, 20]. Table 3 shows SOMI are implemented for imputation with % error classification compared to mean, modus, median and hot desk algorithm [35, 36, 37], respectively. When the level of the missing lower, the better the accuracy imputation. When the rate of missing grows higher, the algorithm undetermined the correct values in all cases and the higher relative error. In all datasets, mean and median algorithm showed similar performance. Modus and hot desk imputation performs better than mean and median in most of these cases. Mean imputation finds the average of certain variables with complete data to replace each row of features that correspond to their average position. Modus imputation is used for a set of discrete data and shows better results than mean imputation, especially for sparse data [35]. Meanwhile, median imputation is used to find the middle value of all known data values to replace the other values. Hot Deck is an improvement of missing data imputation using mean/average, modus or median mode to provide better results compared to deleting incomplete data or when compared to imputation methods with constant values [3].

Table 2. Description of dataset Used in the experiments

Data	**Type**	**Feature**	**Numb. sample**	**Numb. of Class**
Glass	Real	10	214	7
Heart Disease	Mixed	14	303	2
WDBC	Real	32	569	2
Sonar	Real	60	208	2
Record Linkage Comparison Patterns	Discrete	12	5749132	2

Table 3. Imputation results (MSE %) for dataset from different methods

Dataset	**SOMI**	**Mean**	**Mod.**	**Med.**	**Hot Desk**
Glass	1.23	4.28	1.39	4.58	1.38
Heart Disease	3.61	2.34	1.82	2.98	2.01
WDBC	3.45	3.89	3.89	3.13	3.24
Sonar	6.23	6.53	5.47	4.21	2.42
Record Linkage Comparison Pat.	4.89	10.23	9.24	5.45	5.56

The general process for selection and classification of missing values was done by the hybrid GANB method. Finally, GA was used to analyze the chromosome subset to extract the selected *n* features. The accuracy measurement scenario for training was carried out with missing data at 5%, 10%, 20%, 30% and 40%. SOMI determined the parameters and the initial network weights were generated randomly [12]. In this study, the experiments were done with a population size of 100. Table 1 shows an illustration of *n* chromosomes as the initial population. Then, we used two-point crossover for applying a Gaussian mutation operation to the probability of each line of descent from crossover and mutation. Each bit of each chromosome is a representation of one feature. Bit 1 and bit 0 represent the presence or absence of features in an individual. For example, the GA selection method yielded $x_2, x_3, x_4, x_6, x_{8,} x_{10}$ and $x_1, x_5, x_7, x_9, x_{11}$ as selected features. The next step was a partial test to find out which variables significantly influence the selection results. Table 4, the maximum number of generation was varied at 10, 50, 100, 500 and 1000, where each run was executed five times. The results show that the greater the number of iterations the higher the yield stability produced and vice versa. The selection process only draws the required attributes after discretizing the data. Then, accuracy measurement was done for each scenario using cross validation. We have 10 rounds of cross validation. In each round, instances belonging to one of the subjects are used as test data, while instances of the remaining 9 parts are used as training data [24, 27]. Probabilities Mutation ($p_{m)}$ and Crossover (p_c) are set empirically to limit the number of features selected, if a better set is produced to avoid overfitting caused by selecting unnecessary feature sets [30, 34].

Tables 5, 6, 7, 8 and 9 show that the classification accuracy of missing values by SOMI-GANB was better than that of the other algorithms because heuristic search was carried out before imputation. The aim of our experiment is to compare optimizing terms of their ability to choose relevant features GA with without imputation namely SOM, NB and SOMI-NN. GANB selected feature set in the NB classification, and to assess its quality quantitatively, we aim to classify subjects based on the completeness of data after imputation process and using the selected features to improve the accuracy of classification [30, 33].

The accuracy of imputation with the hybrid method increased by 1%, showing better performance than the three other imputation models and without imputation. Using GA classifiers such as NB, k-NN and NN on the dataset repository produced significant differences in classification accuracy [22].

Table 4. GA-based iteration of the complete dataset

Data	**GA/NB (generations)**				**NB**
	50	**100**	**500**	**1000**	**All features**
Glass	82.72	80.78	85.78	85.76	78.98
Heart Disease	80.56	78.90	84.76	84.76	77.98
WDBC	78.85	74.35	80.78	82.08	78.87
Sonar	75.78	77.42	79.56	79.56	74.86
Record Linkage Comp Patterns	82.98	87.42	86.56	86.56	83.67

Table 5. Classification accuracy (%) of hybrid imputation method for Glass dataset

Missing rate	**NB**	**SOM**	**GA-NB**	**SOMI-NN**	**SOMI-GANB**
5%	82.38	85.17	86.75	87.777	91.77
10%	81.13	84.06	86.65	88.900	92.92
20%	80.04	83.65	85.98	87.786	93.79
30%	80.00	83.00	85.90	85.986	91.88
40%	79.87	83.81	85.84	84.678	92.98

Table 6. Classification accuracy (%) of hybrid imputation method for Heart Disease dataset

Missing rate	**NB**	**SOM**	**GA-NB**	**SOMI-NN**	**SOMI-GANB**
5%	83.79	82.99	87.88	86.76	91.76
10%	81.91	81.98	86.95	85.98	92.92
20%	81.00	80.79	83.08	82.76	88.77
30%	76.98	80.10	80.89	80.99	84.87
40%	73.88	81.76	72.89	74.67	82.98

Table 7. Classification accuracy (%) of hybrid imputation method for WDBC dataset

Missing rate	**NB**	**SOM**	**GA-NB**	**SOMI-NN**	**SOMI-GANB**
5%	73.87	83.07	85.78	83.77	90.11
10%	80.81	82.77	81.98	85.90	92.92
20%	76.79	80.66	75.01	83.78	90.78
30%	73.01	78.06	76.89	81.98	88.88
40%	70.88	76.87	72.87	73.67	82.98

Table 8. Classification Accuracy (%) of hybrid imputation method for Sonar dataset

Missing rate	**NB**	**SOM**	**GA-NB**	**SOMI-NN**	**SOMI-GANB**
5%	80.07	74.34	82.21	78.62	83.13
10%	81.28	73.23	85.11	75.12	74.65
20%	80.24	75.50	79.24	73.56	72.96
30%	79.21	71.93	78.95	74.78	69.34
40%	75.23	64.91	69.47	67.80	57.67

Table 9. Classification Accuracy (%) of hybrid imputation method for Record Linkage Comparison Patterns dataset

Missing rate	NB	SOM	GA-NB	SOMI-NN	SOMI-GANB
5%	85.87	84.87	85.23	88.65	93.65
10%	82.98	83.98	85.98	85.98	94.87
20%	80.54	85.56	79.82	83.65	92.78
30%	73.65	71.98	78.45	84.76	89.98
40%	70.56	64.80	77.76	77.87	87.98

The hybrid method showed better classification accuracy with large-dimension datasets, with more than 90% classification accuracy at missing rate below 30%. Each classification result showed that a low missing rate yielded better imputation accuracy. However, the higher the rate of development of the missing data, the more likely it was that the algorithm could not determine the right value in all cases, so the error became relatively higher [26]. The experimental results showed that the number of selected features was almost 50% of all features. However, when less features were used the accuracy was not lower compared to the use of all features in the detection process. In general, the results showed that using fewer features in the classification process leads to higher accuracy than when all the features are used in the learning process [24]. Higher accuracy with fewer features is achieved because not all features obtained are important for learning. Therefore, a feature selection process is needed to get the best feature combination for the learning process. GANB classification accuracy was higher than when GA was performed before imputation.

SOMI allows processing of data under partially missing feature and class conditions [13, 20]. When the membership of training data for a particular class is not known at all, it has a vague nature. SOMI will identify the most important features of the training dataset by grouping SOMI nodes to produce output as input into the next process, namely NN and NB classification. Hybrid SOMI with NN [36] applied an adequate representation of the high dimensional input space through the learning process into a lower dimension. The results SOMI-NN learning for Glass Heart Disease, WDBC and Record Linkage Comparison Patterns dataset of 10-cross validations have been produced an accuracy up to 85% compared with NB, SOM, GANB. So, the accuracy of Sonar Dataset results is minimal, because it is influenced by the number of high dimensional features, resulting in inaccurate features especially those without feature optimization compared to those using feature optimization in the classification. Table 6 show SOMI-GANB and GANB without imputation is superior to mixed data [13, 30, 32] because NB has flexibility based on probability for variable data, while NN requires transformation [35]. This study uses the SOMI-GANB hybrid model which is expected to be able to group individual attribute values in the dataset as an imputation and reach the optimal number of clusters with high classification accuracy. The results are compared to the classification obtained without imputation and NN classification. The proposed SOMI-GANB method not only has superior grouping achievements than the method considered, but also achieves better classification accuracy. Because attribute conditions are very influential by increasing the number of conditional attributes, the way to complete missing attributes and the selection of appropriate features [28, 32]. SOMI depends on the conditions of a number of well-chosen clusters for conditional attributes. When the number of clusters in the decision attribute is optimal, the classification associated with each cluster is also close to optimal. In other words, GA provides better grouping and classification performance for complex datasets especially for mixed attributes. The proposed merger of SOMI and GA provides a practical way to optimize the number of cluster attribute values and NB classification accuracy when complex real-world datasets are applied. We propose SOMI hybrid method with GANB has included replacement values through chromosome in the GA to find value in a space-optimized solution to select a feature, after SOMI to replace missing values have been done [29]. The GA algorithm can be applied more complete search with better opportunities to find the optimal solution [30, 31, 34]. Whereas, the Bayesian principle helps in this process by efficiently using covariate values to be used for analysis [30, 33].

6. Conclusion

The hybrid SOMI method combined with GANB performs better than other imputation methods on most data sets with heterogeneous attribute values. The feature selection hybrid algorithm produces an effective and less feature set, as well as the weakest feature enhancements. Since, the NB classification results in higher accuracy in several heterogeneous datasets. GA worked well in finding optimal feature values, achieving imputation results with an error rate < 10%. The population consisted of a collection of individuals in each generation that represented the selected features. Then, the features were extracted based on an exact number of iterations for naïve Bayes classification. The accuracy measurement calculated the performance of naïve Bayes

classification according to the prediction of the exact number of iterations. The system SOMI was able to predict up to 90.00% accuracy with the number of iterations at 1000, i.e. 15% higher than the old system.

Further research, it is necessary to optimize the improvement of feature weights in the imputation process with SOMI by combining feature selection and weighting with GA. In addition, synchronizing the weights according to the learning process is expected to maintain accuracy. The development of a new hybrid genetic algorithm suitable for the model is the development of binary chromosomes mixed with real numbers for the selection and optimization process.

References

[1] W. Shahzad, Q. Rehman, and E. Ahmed, "Missing Data Imputation Using Genetic Algorithm for Supervised Learning", *International Journal of Advanced Computer Science and Applications*, Vol.8, No.3, pp.438-445, 2017.

[2] R.D. Priya and Sivaraj, "Imputation of Discrete and Continuous Missing Values in Large Datasets Using Bayesian Based Ant Colony Optimization", *Arabian Journal for Science and Engineering*, Vol.41, No.12, pp.4981–4993, 2016.

[3] Y. Zhang, "Using Multiple Imputation to Address Missing Values of Hierarchical Data", *Journal of Modern Applied Statistical Methods*, Vol. 16, No.1, pp.744-752, 2017.

[4] M.N.S. Zainudin, M.N. Sulaiman, N. Mustapha, T. Perumal, A. Shahrel, A. Nazri, R.Mohamed, and S.A. Manaf, "Feature Selection Optimization Using Hybrid Relief-f with Self Adaptive Differential Evolution", *International Journal of Intelligent Engineering and Systems*, Vol.10, No.2, pp.21-29, 2017.

[5] R. Armina, A.M. Zain, N.A. Ali, and R. Sallehuddin, "A Review On Missing Value Estimation Using Imputation Algorithm", In: *IOP Conf. Series on Journal of Physics*, Vol. 892, pp.1-12, 2017.

[6] M. Borrottia, G. Minervinid, D.D. Lucreziae, and I. Poli, "Naïve Bayes Ant Colony Optimization for Designing High Dimensional Experiments", *Applied Soft Computing*, Vol. 49, pp.259–268, 2016.

[7] P.J.G. Laencina, J.L.S. Gómez, and A.R.F. Vidal, "Pattern Classification with Missing Data: a Review", *Neural Computing and Applications*, Vol. 19, No.2, pp.263–282, 2010.

[8] S.Amiri, B.S. Clarke, J.L. Clarke, and H. Koepke, "A General Hybrid Clustering Technique", *Journal of Computational and Graphical Statistics*, Vol.28, No.3, pp.540-551, 2019.

[9] I.B. Aydilek and A. Arslan, "A Novel Hybrid Approach to Estimating Missing Values in Databases Using K-Nearest Neighbors and Neural Networks", *International Journal of Innovative Computing*, Vol.8, No.7, pp.4705–4717, 2012.

[10] S. Zhang, J. Zhang, X. Zhu, Y. Qin, and C. Zhang, "Missing Value Imputation Based on Data Clustering", *Transactions on Computational Science I, LNCS 4750*, pp.128–138, 2008.

[11] J. Bektaş, T. Ibrikçi, and I.T. Ozcan, "The Impact of Imputation Procedures with Machine Learning Methods on The Performance of Classifiers: An Application to Coronary Artery Disease Data Including Missing Values", *Biomedical Research*, Vol.29, No.13, pp.2780-2785, 2018.

[12] J. Bektaş, T. Ibrikçi, and I.T. Özcan, "The impact of Imputation Procedures with Machine Learning Methods on The Performance of Classifiers: An Application to Coronary Artery Disease Data Including Missing Values", *Biomedical Research*, Vol.29, No.13, pp.2780-2785, 2018.

[13] B. K. Khotimah, Miswanto, and H. Suprajitno, "Adaptive SOMMI (Self Organizing Map Multiple Imputation) Base on Variation Weight for Incomplete Data", In: *Proc. of International Conference on Sustainable Information Engineering and Technology (SIET)*, pp. 82-87, 2018.

[14] P. M. Kiran and K. Patil, "Deep Learning Based Weighted SOM to Forecast Weather and Crop Prediction for Agriculture Application", *International Journal of Intelligent Engineering and Systems*, Vol.11, No.4, pp. 167-176, 2018.

[15] C. Del Coso, D. Fustes, C. Dafonte, F.J. Novoa, J.M.R. Pedreira, and B. Arcay, "Mixing Numerical and Categorical Data in a Self-Organizing Map by Means of Frequency Neurons", *Applied Soft Computing*, Vol.36, pp.246–254, 2015.

[16] C.C. Hsu, Y.P. Huang, and K.W. Chang, "Extended Naive Bayes Classifier for Mixed Data", *Expert Systems with Applications*, Vol.35, No.3, pp. 1080–1083, 2008.

[17] O. Addin, S. M. Sapuan, E. Mahdi, and M. Othman, "A Naive-Bayes Classifier for Damage

Detection in Engineering Materials", *Materials and Design*, Vol.39, No.12, pp.2379-2386, 2007.

[18] P. Domingos and M. Pazzani, "On The Optimality of the Simple Bayesian Classifier Under Zero-One Loss", *Machine Learning*, Vo.29, pp.103–130, 1997.

[19] U.N. Dulhare, "Prediction System for Heart Disease using Naive Bayes and Particle Swarm Optimization", *Biomedical Research*, Vol.29, No.12, pp.2646-2649, 2018.

[20] A. Sorjamaa, B. Maillet, P. Merlin, and A. Lendasse, "SOM+EOF for Finding Missing Values", In: *Proc. of the 15th European Symposium on Artificial Neural Networks*, pp.115-120, 2017.

[21] N. Fytilis and D.M Rizzo, "Coupling Self-Organizing Maps with a Naıve Bayesian Classifier: Stream Classification Studies Using Multiple Assessment Data", *Water Resources Research*, Vol.49, pp.7747–7762, 2013.

[22] K. Homsapaya and O. Sornil, "Improving Floating Search Feature Selection Using Genetic Algorithm", *J. ICT Res. Appl.*, Vol. 11, No.3, pp.299-317, 2017.

[23] L.M. Ibrahim, D.T. Basheer, and M.S. Mahmod, "A Comparison Study for Intrusion Database (Kdd99, Nsl-Kdd) Based On Self Organization Map (SOM) Artificial Neural Network", *Journal of Engineering Science and Technology*, Vol. 8, No.1, pp.107-119, 2013.

[24] Y.L. Wu, C.Y. Tang, M.K. Hor, and P.F. Wu, "Feature Selection Using Genetic Algorithm and Cluster Validation", *Expert Systems with Applications*, Vol.38, No.3, pp.2727-2732, 2011.

[25] V. Bachu and J. Anuradha, "A Review of Feature Selection and Its Methods", *Cybernetics and Information Technologies*, Vol.19, No. 1, pp. 3-26, 2019.

[26] J. Cervantes, X. Li, and W. Yu, "Using Genetic Algorithm to Improve Classication Accuracy on Imbalanced Data", In: *Proc. of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp.2659-2664, 2013.

[27] A. Siradjuddin, R. Septasurya, M. K. Sophan, N. Ifada, and A. Muntasa, "Feature Selection with Genetic Algorithm for Alcoholic Detection using Electroencephalogram", *International Conference on Sustainable Information Engineering and Technology* (SIET), pp.230-234, 2017.

[28] J. Pedro, G. Laencina, J. Luis, S. Gomez, R. Anıbal, and F. Vidal, "Pattern Classification with Missing Data: a Review", *Neural Comput. & Applic.*, Vol.19, Vol.2, pp.263–282, 2010.

[29] R. Malhotra, N. Singh, and Y. Singh, "Genetic Algorithms: Concepts, Design for Optimization of Process Controllers", *Computer and Information Science*, Vol.4, No.2, pp.39-54, 2011.

[30] L.G.P. Suardani, I.M.A. Bhaskara, and M. Sudarma, "Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes Classification for Home Improvement Recipients", *International Journal of Engineering and Emerging Technology*, Vol.3, No.1, pp.66-70, 2018.

[31] M. Pei, E.D. Goodman, F. William, and E.D. Goodman, "Feature Extraction Using Genetic Algorithms", In: *Proc. of International Symp. on Intelligent Data Engineering and Learning'98 (IDEAL'98)*, 1998.

[32] L. Jie and Song Bo, "Naive Bayesian Classifier Based on Genetic Simulated Annealing Algorithm", In: *Proc. of International Conference on Power Electronics and Engineering Applic. on Procedia Engineering*, Vol. 23, pp. 504 – 509, 2019, 2011

[33] M.H.B.M. Adnan, W. Husain, and N.A.R. Ashid, "A Hybrid Approach Using Naïve Bayes and Genetic Algorithm for Childhood Obesity Prediction", In: *Proc. of International Conference on Computer & Information Science (ICCIS)*, pp.281-285, 2012.

[34] S. Chormunge and S. Jena, "Efficient Feature Subset Selection Algorithm for High Dimensional Data", *International Journal of Electrical and Computer Engineering* (IJECE), Vol. 6, No. 4, pp.1880-1888, 2016.

[35] R. J. Little and D.B. Rubin, Statistical analysis with missing data, *Second Edition. John Wiley and Sons*, New York, 2002.

[36] S. K. Gnana and S. N. Deepa, "An Intelligent Hybrid Neural Network Model", In: *Proc. of Renewable Energy Systems, International Conference on Computer Technology and Science (ICCTS 2012) IPCSIT*, Vol. 47, pp.181-184, 2012.

[37] T. Kim, W. Ko, and J. Kim, "Analysis and Impact Evaluation of Missing Data Imputation in Day-ahead PV Generation Forecasting", *Applied Science*, Vol.9, No. 204, pp.1-18, 2019.