International Journal of GEOMATE, Oct., 2019 Vol.17, Issue 62, pp. 195 - 202
ISSN: 2186-2982 (P), 2186-2990 (O), Japan, DOI: https://doi.org/10.21660/2019.62.71789
Special Issue on Science, Engineering & Environment

A HYBRID SELF ORGANIZING MAP IMPUTATION (SOMI) WITH NAÏVE BAYES FOR IMPUTATION MISSING DATA CLASSIFICATION

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

[1]Faculty of Science and Technology, University of Airlangga, Surabaya, East Java, Indonesia;
[1]Faculty of Engineering, University of Trunojoyo Madura, East Java, Indonesia;
[2,3]Department of Mathematics, University of Airlangga, Surabaya, East Java, Indonesia;

*Corresponding Author, Received: 18 Feb 2019, Revised: 21 March 2019, Accepted: 08 April 2019

ABSTRACT: This study proposes hybrid SOMI (Self Organizing Map Imputation) and Naïve Bayes (NB) model on data, that contain missing values to improve the performance of the Naïve Bayes Imputation (NBI) it has weaknesses for missing categories n ≤ 1. This new hybrid model, using imputation approach based on SOMI is used for prepossessing and NB classification for the classification process in multivariate data, so that it can improve performance. SOMI measurements use an average error with self-organizing feature map. The multivariate attribute is converted to numeric attributes to establish data uniformity. The SOMI learning results have used weight variations by combining the mechanism of distance hierarchical value representation with a new scheme to overcome mixed types. Hybrid SOMINB is used to classify mixed data to correct misclassification. The model has advantages because it can update weights with the probability of each attribute. Attribute values have produced a set of probabilities for each cluster using the Naïve Bayes group. Outputs of the SOMI Method are used as learning machines to produce training data for the target class to be used in Naive Bayes machine learning. The results of this study used all missing scenarios at a random mechanism and various missing percentages. The results of the hybrid SOMINB model showed more results with an accuracy rate of 90.00% with other imputation analysis. Experimental results present that the proposed produces higher accuracy than general estimating values which established missing value treatment methods.

Keywords: Missing data; Naïve Bayes; SOMI, Imputation; Hybrid Model; Classification

1. INTRODUCTION

Many industrial and medical datasets contain missing value due to errors of data entry procedures, measurement errors and unfamiliarity respondents in answering [1-2]. The technique to overcome missing value with the simple method only fits the data relatively small [3]. Research development in recent years uses machine learning (ML) algorithms for imputation methods [4]. Research for other missing data is performed by Madhu G. and Nagachandrika G., with they compare imputation techniques with statistical techniques, that is mean, hot deck, multiple imputations to imputation [5]. A method based on machine learning techniques for missing value namely Multilayer Perceptron (MLP), Self Organizing Map (SOM), and k-NN on breast cancer data. Some applications of classification machine learning on missing data are k-Nearest Network Imputation (k-NNI), Support Vector Machine Imputation (SVMI), Neural Network Imputation (NNI), and Decision Tree Imputation (DTI) [19-23] [6-10]. k-NNI was measured for the closeness between data on missing data depending on *k* to measure the data similarity level, if the higher the *k* value will result in bias in classification [8]. Other imputation method, NNI can be used on any data, whether continuous or category. Although the data need for preprocessing data prior training so that the resulting calculation will be longer [11]. Naive Bayes Classification (NBC) for handling missing data need appropriate replacement value to maintain the method performance. Missing data at multivariate if there are mixed values either discrete, continuous, and category will require the conversion process to be numerical value [12]. NBC to handle missing data can work with the condition it requires imputation process firstly to replace value part whose attribute missed so it is called Naive Bayes Imputation (NBI). The use of NBC has advantages with a low variance to reduce the high bias effect because of the assumption of strong features independence [13]. When the assumption of NBC attributes independence is violated, so the classification performance of the NBC class is very poor and the classification accuracy decreases. The development imputation use to improve the model quality, with a number of approaches attempting to reduce the assumption of independent NB grouping problems, which is namely using the hybrid k-means Imputation (KMI) model with NB [14]. The KMINB model uses the test dataset attribute with posterior probability with observation partition to the cluster *C* to produce variant weights as the value of imputation [15]. Many researchs have been carried out on hybrid models to overcome missing values. Fatemeh Ahmadi Bakhsh et al., also propose a clustering-based method for preprocessing the data to replace the missing weight for the complete data. Cluster is generated from a non-labeled data set will group according to requirements [14].

Hybrid models of the SOM and NB algorithms are three phases: pre-processing and feature selection and NB classification in optimization groupings. The grouping process will be optimized based on probability feature by applying to preprocess for SOM grouping. The classification and imputation phases using centroids remain as variants in each appropriate cluster [16]. Since, imputation will fill the process with the NBI classification for supervised income for grouping hidden class variables for feature variables discrete or continuous [22].

This research proposes a new hybrid model by combining SOMI with NB where the weights result between each data sample and cluster center that have one-dimensional distance feature that is used for the imputation process. Clustering technique on missing data is done to overcome errors in the classification of the same class, where as the replacement weight is higher or lower than the actual value. The hybrid method provides maximum accuracy and has resistance to missing data. If using NBI has weaknesses that provide the highest accuracy value, but this method cannot measure heterogeneous levels in groups and homogeneous levels between groups.

The research was conducted as follows: Sec. 1 presents introduced the SOMI technique to take over data of missing values. Sec. 2 presents a theoretical background about missing values that followed by the Hybrid SOMINB proposed. In sec. 3, the evaluated criteria for measuring performance. In sec. 4, The research methodology to describe the course of study. In sec.5, Results and discussion. Finally, this section explains conclusions and so on.

2. SOM (SELF ORGANIZING MAP)

Basically, SOM do data collection mapping that has dimension *d*, which is a series of arrays containing both discrete and continuous data on mapping dimension [17]. The SOM network has two layers, namely the input layer and the output layer where each neuron in the output layer represents the class of input provided. Each $\Re^d$ node in given the weight vector $\overline{m}_i$, each data input unit has vector input with $\overline{m}_i \in \Re^d$. Then, the comparison of distances between $\overset{v}{x}$ and tested to get mapping results. Training data matrix with dimensions of sample *N* sample *d*, then $\Re^d$ mapping for N^2 can be stated as f: $\Re^d \rightarrow N^2$ and vector mapping functions as weight vectors to produce target groups [18]. The selection of the best neurons is based on the time to produce similarity matching. The usage of neurons is the smallest distance between vectorthe $\overset{v}{x}$ and weight vector $\overline{m}_i$ as the initial weight for all neurons. Calculation of the smallest distance using the $\|\bar{x}t - \overline{m}_i\| = min \sum_{i=1}^{n} \|\bar{x} - \overline{m}_i\|$ the following with Euclidean distance with using with c= argmin $\|\bar{x}t - \overline{m}_i\|$ by using i=1,2,…,n,c is the best index of neurons. So, $\left\|\overline{\underset{v}{x}}(t) - \underset{v\,c}{m}\right\|$ is the distance between input vectors $\overset{v}{x}$ and weight vectors can be calculated with $\|\bar{x}t - \overline{m}_i\| = \sum_{i=1}^{n} (\bar{x} - \overline{m}_i)^2$. The basic idea when updating the reference vector, all data calculated on the update rules is [19]:

$$m_i = \frac{\sum_n h_{ni} x_n}{\sum_j h_{ni}} \qquad (1)$$

So, with n indexes running according to vector data if $h_{ni} > 0$ then it will be updated vector references using all data points to various environmental functions.

2.1 SOMI (Self Organizing Map (Imputation)

Each SOMI map has a different number of nodes for estimation of missing data value in different databases. SOMI provides the final estimate of the missing value $\hat{y}$ based on linear combinations of weights generated. The weight is $w_{nk} = 1 \forall x_{nk}$ for complete and $w_{nk} = w \leq 1 \forall x_{nk}$ for incomplete data. The weight $w \leq 1$ was used to replace the missing value from the final process of SOMI learning. When each observational data containing missing values, the value of each weight of best BMU from observations will be used to replace the missing value. The SOMI update process is carried out with continuous weight changes until experience convergent stated as [20]:

$$m_i = \frac{\sum_n h_{ni} w_n . x_n}{\sum_j h_{ni} w_n} \qquad (2)$$

after studying the SOM algorithm it has converged, the results of several clusters that contain weights and data have been grouped. Process of missing values was filled in the dataset by using vector coordinates based on the code of each BMU to replace the missing value by:

$$X_{(M_x)}(x) = X_{(M_x)}\left(m_{BMU(x)}\right), \qquad (3)$$

Variable weights $X_{(M_x)}(x)$.) replace the missing value at M from sample x. The replacement of missing data is carried out for each data sample with the final weight of SOMI learning [21]. Process of code filling in the training data set requires party value x_i class value c_k, and attribute values w_{ij}. In addition, to facilitate filling by using numerical data attributes. The processing procedure for filling in the missing value that suitable proposed method. It will become the object estimation index x_{im} to i on the winning node (best match) of the unit c_{ij}, that from the order of variables to j by process SOMI is determined for *i*-th $X_{ir} = (c_i, w_j)$ when there are criteria *r* containing missa ing value for:

$$1 \le w_{ij} \le p \text{ , if } w_{ij} = x_{im}(i,m) \in N_m, \qquad (4)$$

Values of learning SOMI is the form of weight at each data in the same class with $w_{ij} : (i, j) \in M_0$. While the replacement weight of missing data will be resulted the new database after experiencing of the imputation process stated:

$$v_{ij} = x_{im}(i,p) \in N_p, \text{and } \{v_{ij} : (i,j) \in N_p\} \qquad (5)$$

A prototype of a weight vector occur all nodes with natural number is the index of nodes. Then, *l* is the number of nodes that make up the competitive layer. Selectio $m_l \in \{1, 2, ..., l\}$ m_c of as the winning node is the best matching unit for data input *x* which still contains missing values state the following formula:

$$m_c = \arg\min \left\| m_{\{v\backslash k\}l} - x_{\{v\backslash k\}l} \right\| \qquad (6)$$

with *k* is the attribute set with a missing value. The weight vector for complete data is defined as $m_{\{v\backslash k\}l}$ and $x_{\{v\backslash k\}l}$. This weight vector from the winning node for the data that includes the missing value is specified. Then, the estimated value is obtained and replaced with missing values. SOM based missing value imputation method reflects the distribution and features of missing values and complete data. Measurement of the overall error value in the *i* test in ensemble stated [37]:

$$\hat{m}_i^{bmu} = \frac{1}{E}\sum_{j=1}^{E} m_{c_{ij}} \qquad (7)$$

The smallest error value involves better prediction results, and vice versa.

2.2 Hybrid SOMI and Naïve Bayes (SOMINB) Model

A final result is a group of data and weight vectors in the win *i* node what used to estimate missing value in the object. Weight vectors are used to fill missing attribute data x_i^m. Each variable from input data x_i follow the normal density distribution for each neuron *n*, because every neuron in the output grid will have weight with the same previous probability i.e. $p_n = 1/n$. The mathematic definition is expressed in equation [13]:

$$p(x_i \mid n) = \frac{1}{\left(2\pi\sigma_{in}^2\right)^{\frac{1}{2}}} \exp\left(-\frac{\left(x_i - w_{in}\right)^2}{2\sigma_{in}^2}\right) \qquad (8)$$

with *p(x,X)* is the empirical probability density estimation at *x* in space X data, w_{in} is the average weight, and σ_{in} is a standard deviation. Data is calculated using normal density distribution for certain neurons. Then, NB for the missing value imputation, with the class variable given $\theta_k = \{w, \sigma_{ip}\}$ with $w_{ip} = \{w_{11}, \dots w_{ip}\}$ and $\sigma_{ip} = \{\sigma_{1p}, \sigma_{2p} \dots . \sigma_{ip}\}$ determined by [16]:

$$p\left(w_{ip} \mid k, \theta_k\right) = \prod_{i=1}^{n} p\left(w_{\ ip} \mid k\right) \qquad (9)$$

The probability distribution p of all x inputs for all test data are given formula:

$$p\left(x \mid \Theta\right) = \frac{1}{K}\sum_{k=1}^{K} p\left(x \mid k, \theta_k\right) \qquad (10)$$

With $\Theta = (\theta_1, \theta_2 ..., \theta_k)$ for the mixed model is converted to discrete value form by making an equivalent variable value. NBC uses the Gaussian i function with vector point k which can be done by finding the class value by maximizing the probabilistic value to determine posterior class maximization on a number of attributes. The attribute column at the missing value, so the missing value is replaced by the weight on SOMI x_m , the formula for the class value is the same $X_i = x_1^r, x_2^r, x_2^m, ..., x_i^r$ with,

$$C(X_i / k) = \arg\max_k P\left(C = k \mid X_i\right) \qquad (11)$$

Where *P* is recorded data from cluster j which belongs to class *i*; m is the number of clusters; nj is the amount of data in cluster j and n is the sum of all data [22].

3. EVALUATION CRITERIA

Evaluation criteria have contained elements of *TP* (True Positive), *FN* (False Negative), *FP* (False Positive), and *TN* (True Negative). The measurement has used Precision (*PR*), Recall and F-value. *PR* is a measurement of data to predict positivity with actual positive results. The recall is the proportion of data grouped according to the label so that the classifier gets better by having a high recall value. Then, F-value is the variance value of Mean Square Error (MSE) between the divided mean of within by group variance [39].

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

$$FV = \frac{\left(1+\beta^2\right) xRecall\ xPrecision}{\beta^2 x\left(Recall+Precision\right)} \quad (14)$$

Furthermore, this stage the Cross Validation method is also carried out in testing the resulting model. In Cross Validation, the dataset used is divided into sections (*n*) [22].

4. RESEARCH METHODOLOGY

Imputation of missing data with SOMI and NB classification is testing the correlation data to avoid dependencies features that greatly affect performance. Missing data is done by randomly removing some valueson several variables in each dataset to get missing values of 5%, 10%, 15%, 20% and 30% of total data. The data is used that contains missing values to be included SOMI model for estimating weights. The appropriate weight SOMI for the imputation process suitable the group will be targeted. Furthermore, learning is done on the same data in Naive Bayes using the target class as a result of evaluating SOMI learning. NB learning is done on complete data from the results of missing data imputation with the final weight of SOMI learning. Data normalization uses Z-Score transformation because in Naïve Bayes classification requires real data if the data is mixed. To predict x labels in each class C_j, then the usage is the greatest opportunity value. Finally, they will be doing an error calculation and method analysis.

The research perform a series of computational experiments on the set of data taken from the UCI Machine Learning Warehouse for assignment and classification hybrid SOMINB model. The data set shows a number of observations with a number of features with data mix consisting of continuous variables and categories as well as continuous and nominal variables in each set of data is given in Table 1.

Table 1. Data sets for classification experiments

Data Set	N	Cont.	Nom.	Cat.	C
Heart	270	6	0	7	2
Hepatitis	80	6	0	13	2
Australian Credit	690	6	0	8	2
Horse Colic	368	7	15	0	2

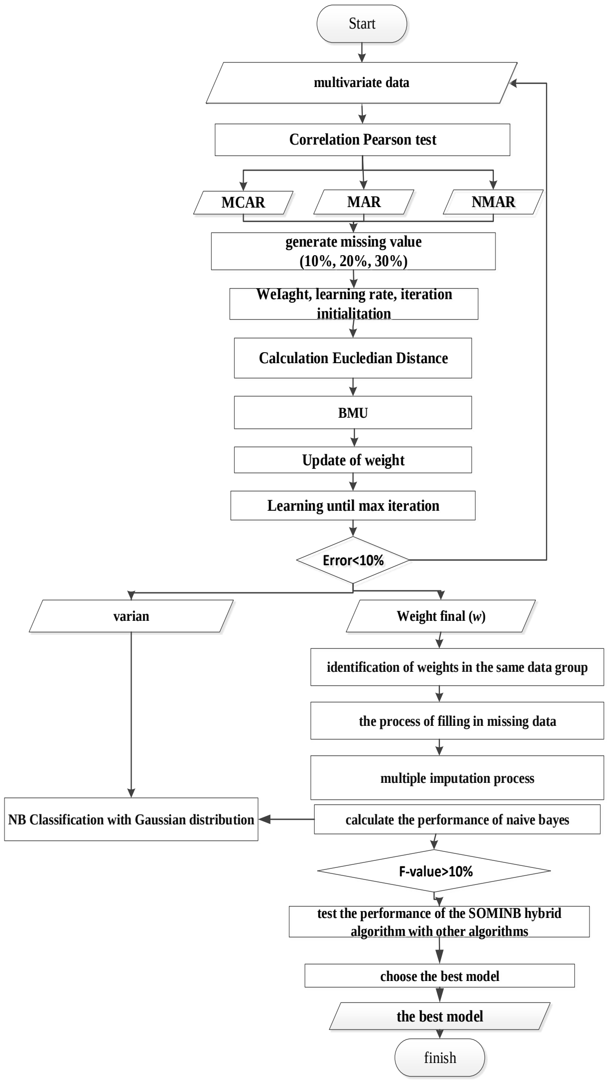
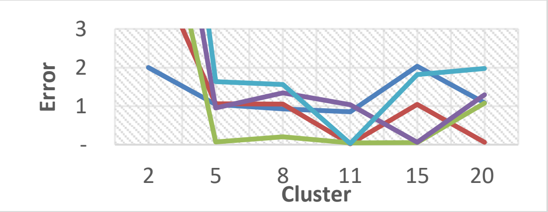

Start
multivariate data
Correlation Pearson test
MCAR
MAR
NMAR
generate missing value (10%, 20%, 30%)
WeIaght, learning rate, iteration initialitation
Calculation Eucledian Distance
BMU
Update of weight
Learning until max iteration
Error<10%
varian
Weight final (w)
identification of weights in the same data group
the process of filling in missing data
multiple imputation process
NB Classification with Gaussian distribution
calculate the performance of naive bayes
F-value>10%
test the performance of the SOMINB hybrid algorithm with other algorithms
choose the best model
the best model
finish

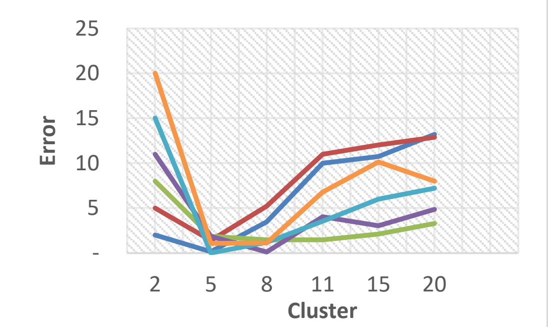
Fig. 1 Flowchart of proses imputation with hybrid SOMINB models

5. RESULTS AND DISCUSSION

Experiments were carried out on machines using Pentium Intel (R) Core (TM) i5 processors with missing data taken 5%, 10%, 15%, 20%, and 30%. Experimental results shown in Tables (3-7) show to complete data sets of mixed scales which transformed first by using z scores. So, that no variable dominates Euclidean values and the potential uses probabilities during the mapping process. Setting parameters in prepossessing are used in experiments with 1.000 iterations. SOMI has been updated by determining the number ensemble *n* with 12 for the number of nodes. SOMI is applied to get imputation weights in each training class and to get a weighted vector optimized to represent the appropriate class. The numbers of weighting vectors are much smaller than the original sample in the training class. Furthermore, SOMI learning optimized in each vector use Euclidean distances between input vectors and weight vector $d(X_i, W_i)$, with elements $w_{1j} = j_{th}$ based on certain neuron number. BMU is best obtained when neuron that has the closest vector matches the vector input data with $d(X_i, W_i)$ is minimum. To maintain accuracy, the new SOMI technique with Naive Bayes replaces missing data based on the weighting of SOMI grouping as a pre-classification component. The experimental results show that each level of learning produces weights to estimate the missing value. Each weight is taken the average value of each learning outcome. For example, a data set of n=3 with group $\Omega = \{\omega_{11}, \omega_{11}, ..., \omega_{n3}\}$ and others. The results of the SOMI ensemble calculation use error on average suitable equation (**7**). They have contained the parameter, with *k* is a cluster, c_k is the middle size, and x is observation data, subscript $x \in c_k$. Since, it is the observation that part of the cluster is centered on particular centroid. The input pattern *x* can be mapped to SOMI at the location *(i,j)* where $\sigma(i, j)$ is the minimum distance to *x*. Competitive SOM learning is adopted, and training algorithms are repeated to achieve maximum results. Thus, BMU whose weight vector is most similar to the input vector is found with the magnitude of the change decreasing over time and with distance from BMU. The ensemble error pattern on each data is shown in Fig. 2(a-b).



(a).



(b).

Fig. 2 (a-b). The relationship shows error towards a positive grouping of data sets

In each plot, the numbers of clusters that run on the horizontal axis, and the error value based on learning in the ensemble. Exactly, they located on vertical line shows the addition of clusters allows an increase in errors. Figure 2 shows the index value with some local minimum values. Because it is very sensitive to initialization learning rate parameters and a number of ensembles randomly selected. To choose the best grouping with a different number of clusters, sharp local minimums in the validity index plot are certainly worth seeing. Because this shows that adding one cluster allows the algorithm for partition the data much better. In the data set of Heart, the error drop point is very sharp at the 11th cluster and has a significant increase in the error limit below 2. While in the Horse Colic data the error lies in cluster 5-8, but after that, it rises with an error exceeding 10. The imputation algorithm experiment is done with reference vector $m_i(t = 0)$, learning rate (0.1,10) and (0.5,20), iterations n= 1000 and Gaussian noise 0.01.

Table 3. summarizes the experiment results with 4 data sets. The result show error analysis which used several input methods. They were not statistically significant with error depending on the width of the working environment at SOMI. In addition, data sets that have used attributes continuously with discrete and combinations of discrete attributes and categories produce insignificant differences. Only if the number of derivation instances per class value is not balanced gives greater error value, the result is better to use continuous attributes. If the data has concluded the mix requires continuous attribute discretization, then it potentially reduces errors in a mix of attributes and categories on an ongoing basis. In experiments, the SOMI grouping algorithm that was trained using the four data shows effective results for imputation. Method SOMI perform better than all the methods, when the missing proportion almost at all proportions. The experiment results using differences in the number of ensembles and learning rates (n, ρ) shows that the more value of *n*. The model results could be reported on Table 6 that SOMI algorithm showed the more superior to the ensemble k =20 and the learning rate is close to $\rho \geq 1$. Base on the experiment results, the observed increase in heart datasets could be attributed to the model proposed is more effective than other datasets. So, error value is lower with an increasing percentage of missing values. Since, the validation technique used is cross validation of 10 times, by dividing the dataset into 10 parts. While data consists of 10 parts, parts of 9 are used as training data and the remaining 1 part is used as test data. Based on the experiment results, performance comparison of imputation without it was carried out to determine the best classification algorithm. Measurement Table 1 is done by examining four datasets UCI machine learning repository (Heart, Diabetes, Australian Credit and Horse Colic). One of the problems in grouping missing values is choosing the right weight to replace missing data with high accuracy performance. Often, because the weight is irrelevant and excessive in value, it affects the classification results. Data processing is very important to assign features to the cluster closest with optimizing centroid clusters are updated by the NB algorithm by testing each of the most suitable replacement values. The model proposed by constant centroid for the cluster influenced by differences in several learning rate parameters, cluster numbers, and ensemble for testing by taking the average value.

Table 3. Results difference in method performance imputation value for mixed attributes

Data	% Missing Value	Error (*E*)							
		(0.1, 10)				(0.5,20)			
		SOM	Mean Imp.	hot deck Imp.	SOM Imp.	SOM	Mean Imp.	hot deck Imp.	SOM Imp.
Heart	5	1,225	5,345	3,343	1977	2.151	4.5 34	2,567	0.934
	10	2,058	6.221	4,256	2,700	1,023	4.1 03	5,138	1,746
	15	5.209	6.984	4,453	3,055	1,145	6,246	5.209	0.055
	20	4,345	7,453	8,234	2,975	1,939	8,682	10,388	0.965
	30	6,563	10,340	15,247	0.817	1,036	12,520	14.126	1,817
Hepatitis	5	3,026	5,520	2,230	0.765	2.151	2,453	4,026	0.934
	10	1,052	6,502	6,502	1956	1,023	4,879	8,012	1,045
	15	1,731	5.342	5.342	1955	1,145	5.2 85	10.785	0.357
	20	2,305	6,321	8.321	1,678	1,939	6. 673	12,378	0.665
	30	1,573	10,011	12,011	0.317	1,775	10. 087	12,643	1,098
Australian Credit	5	0.826	4,414	3,094	2,765	2.165	5.145	10.826	2,934
	10	1,052	6,232	4.130	4,956	3,023	8,198	13,052	3,678
	15	2,731	7.328	6,358	3,955	2,873	6,678	16,731	2,789
	20	3305	10.221	10,276	1,678	2,938	9,245	19.305	3,890
	30	4,573	11,412	10,432	2,512	3,590	18,456	21,573	5,900
Horse Colic	5	1,921	6,121	3,121	2,566	0.165	8071	2,921	1904
	10	1,073	8,002	7,034	2,563	1,023	5,012	4,354	6,678
	15	1,635	9,090	10,090	2,257	4.981	7,452	8.102	8,785
	20	2,392	11,215	12,432	2,678	2.187	14.167	9,192	10,879
	30	2.401	14,906	15,342	3,513	4,544	13,674	12,785	11.903

Table 4. Accuracy percentage of heart data sets experiment method using all methods.

% Miss.	NB			Hot deck+NB			Mean+NB			SOMI+NB		
	Prec.	Rec.	F-value	Prec.	Rec.	F-value	Prec.	Rec.	F-value	Prec.	Rec.	F-value
5	82.56	77.11	74.56	8 5.01	87.12	84.26	87.66	90.10	89.46	88.54	91.92	85.67
10	80.13	78.28	73.32	83.81	85.21	83.42	86.81	88.20	86.40	87.76	90.72	82.78
15	75.67	74.20	70.12	85.20	84.27	80.20	84.23	86.34	87.13	86.08	89.78	80.56
20	7 8.59	66.44	68.2 4	71.57	76.43	78.74	78.16	82.34	80.14	75.56	88.00	78.71
30	76.23	63.20	59.1 6	65.29	73.2 9	69.36	70.15	75.25	80.34	74.45	84.09	76.00

Table 5. Accuracy percentage of diabetes set data experiment method using all methods.

% Miss.	NB			Hot deck+NB			Mean+NB			SOMI+NB		
	Prec.	Rec.	F-value	Prec.	Rec.	F-value	Prec.	Rec.	F-value	Prec.	Rec.	F-value
5	90.60	90.16	91.88	84.56	89.80	88.9 6	90.86	91.34	91.59	92.90	90.57	93.65
10	87.26	89.40	87.56	82.32	87.23	88.40	88.70	89.89	88.72	92.30	87.89	90.69
15	85.73	86.78	88.57	80.12	83.12	8 5.79	88.89	88.49	87.23	89.56	84.03	89.66
20	82.22	84.32	85.30	78.2 4	80.38	84.74	86.78	87.77	85.47	87.18	83.56	88.70
30	80.00	83.19	82.20	67.1 6	75.32	82.36	80.80	83.78	82.65	86.12	79.90	84.20

Table 6. Accuracy percentage of Australian Credit data set experiment method using all methods.

% Miss.	NB			Hot deck+NB			Mean+NB			SOMI+NB		
	Prec.	Rec.	F-value	Prec.	Rec.	F-value	Prec.	Rec.	F-value	Prec.	Rec.	F-value
5	76.89	87.45	90.34	83.56	88.76	89.60	84.85	81.19	88.65	81.59	80.12	90.30
10	87.76	81.20	88.24	80.45	84.45	87.65	81.56	79.24	86.74	80.70	81.56	87.28
15	81.00	80.29	75.17	84.34	83.20	80.71	78.54	78.46	90.34	78.07	78.88	80.36
20	76.90	76.24	78.59	77.50	78.53	78.52	73.98	76.78	89.45	75.16	77.40	78.78
30	75.34	73.68	76.23	75.10	77.20	73.49	70.89	75.65	86.86	74.15	76.29	75.12

Table 7. Accuracy percentage of Horse Colic data set experiment method using all methods.

% Miss.	NB			Hot deck+NB			Mean+NB			SOMI+NB		
	Prec.	Rec.	F-value	Prec.	Rec.	F-value	Prec.	Rec.	F-value	Prec.	Rec.	F-value
5	57.21	79.40	72.77	68.51	80.71	81.78	80.85	82.16	82.17	8 0.12	87.63	89.24
10	77.06	77.26	70.34	70.45	81.15	78.60	78.56	79.45	80.93	80.20	85.45	85.98
15	78.20	70.29	69.18	64.34	80.17	76.54	77.54	80.40	80.37	7 8.54	83.60	84.89
20	73.39	66.01	68.78	67.50	79.56	74.92	73.98	76.56	76.11	75.15	82.43	80.12
30	67.23	65.17	65.24	65.10	78.11	72.41	60.89	75.34	70.65	74.27	80.13	79.10

The experimental results show the difference between the four methods in the Highest SOMI+NB. Exactly, the experiment did not have an insignificant difference (1%). Table 4 shows that the SOMI+NB produces the best Precision, Recall and F-Value which the highest percent precision compared to the other methods. But They are without significant difference of 10% compared to other imputation methods. Table 5 found that the results were higher in Precision, a slight increase in Recall percentage. While a slight difference is 1%, Table (4-7) experiences an increased precision with minimal increase of 57.21% and a maximum increase of 67, 23% using NBI. The hybrid SOMI model with Naïve Bayes has been tested with 4 databases that are commonly used. The quality of SOMI cluster results is proven by the lower error. So, research on the subject has been mostly restricted to comparisons among the proposed models and other approaches to determine the accuracy of the system performance. A new approach model in NB automatic grouping with mixed data by calculating probabilities feature after all data have been filled through pre-processing. Learning outcomes using SOMI methods and Mean indicate recall and F-value above 80%. The classification results formed can be used as a new approach with grouping models that are influenced by the number and diversity of data sets that are owned. quality mapping.

6. CONCLUSIONS

In experiments, the SOMI clustering method provides higher accuracy than conventional imputation methods when estimating missing values that are randomly generated. Clustering by adding the ensemble method can increase the probability of the Naïve Bayes classification, so it can handle missing values randomly. The new hybrid classifier model combines ensemble clustering and Naïve Bayes with classification levels by increasing accuracy. Experiments comparison results of classification by using four sets of data, then the classification SOMINB show F-value is very effective in reducing the number of instances and increase the sensitivity and recall value. A hybrid model of future research issues can be improved in the areas of real data with time complexity analysis that is fast and efficient.

7. ACKNOWLEDGMENTS

The authors thank the promoters who provided comments as feedback to improve of the final manuscript. We also thank the staffs at University of Airlangga Surabaya who helped with the administrative process during the research.

8. REFERENCES

[1] Zahrah S.N., Amin Burhanuddin, Deni S.M. and Ramli N.M., Normal Ratio in Multiple Imputation Based On Bootstrapped Sample for Rainfall Data with Missingness, International Journal of GEOMATE, Vol.13, Aug 2017, Issue 36, pp.131-137.

[2] Little R.J.A. and Rubin D. B., Statistical analysis with missing data, second edition Wiley, 2002.

[3] Bertsimas D., Pawlowski C. and Zhuo Y.D., From Predictive Methods to Missing Data Imputation: An Optimization Approach, Journal of Machine Learning Research, Vol.18, 2018, pp.1-39.

[4] Vaishali H., Umathe and Chaudhary G., A Review on Incomplete Data and Clustering, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6, Issue 2, 2015, pp. 1225-1227.

[5] Madhu G., and Nagachandrika G., A New Paradigm for Development of Data Imputation Approach for Missing Value Estimation, International Journal of Electrical and Computer Engineering (IJECE), Vol. 6, Issue 6, December 2016, pp. 3222 – 3228.

[6] Afza A. J. M.A., Farid D.M. and Rahman C.M., A Hybrid Classifier using Boosting, Clustering, Naïve Bayesian Classifier, World of Computer Science and Information Technology Journal (WCSIT), Vol.1, Issue 3, 2011, pp.105-109.

[7] Buuren S.V., Multiple imputations of discrete and continuous data by fully conditional specification, Statistical Methods in Medical Research, Vol.16, Issue 3, 2007, pp. 219–242.

[8] Priya S. and Thanamani A.S., Comparative Study of Naïve Bayes Classifier and K Nearest Neighbor in Imputation of Missing Values, International Journal of Trend in Research and Development (IJTRD) in National Conference on Digital Transformation – Challenges and Outcomes (ASAT in CS'17) o, Bangalore on 3rd Mar 2017, pp.12-14.

[9] Joenssen D.W., Hot Deck Methods for Imputing Missing Data, Machine Learning and Data Mining in Pattern Recognition, Springer, Vol. 7376, July 2012.

[10] Gimpy, Vohra R. and Minakshi, Estimation of Missing Values Using Decision Tree Approach, International Journal of Computer Science and Information Technologies, Vol. 5, Issue 4, 2014, pp. 5216-5220.

[11] Doquire G., and Verleysen M., An Hybrid Approach to Feature Selection for Mixed Categorical and Continuous Data, Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, 2011, pp. 394-401.

[12] Subramanian U. and Ong H.S., Analysis of the Effect of Clustering the Training Data in Naive Bayes Classifier for Anomaly Network Intrusion Detection, Journal of Advances in Computer Networks, Vol. 2, Issue 1, March 2014, pp. 85-88.

[13] Hsu C.C., Huang Y.P. and Chang K.W., Extended Naive Bayes classifier for mixed data, Expert Systems with Applications, Vol. 35. 2008, 1080–1083.

[14] Bakhsh F.A. and Maghooli K., Missing Data Analysis: A Survey on the Effect of Different K-Means Clustering Algorithms, American Journal of Signal Processing, Vol.4, Issue 3, 2014, pp. 65-70.

[15] Varuna S. and Natesan P., An Integration of K-Means Clustering and Naïve Bayes Classifier for Intrusion Detection, 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), 2015.

[16] Gonzalo A.R.D. and Truong P., NBSOM: The naive Bayes self-organizing map, Neural Comput & Applic, Spinger, Vol. 21, 2012, pp.1319–1330.

[17] Fort J.C., Letrémy P. and Cottrell M., Advantages and Drawbacks of the Batch Kohonen Algorithm, in ESANN, 2002, pp. 223–230.

[18] Ahmad A. and Yusof R., A Modified Kohonen Self-Organizing Map (KSOM) Clustering for Four Categorical Data, Jurnal Teknologi, Vol.78, 2016, pp. 6-13.

[19] Vatanen T., Osmala M., Raiko T., Lagus K., Sysi A.M., Orešič M., Honkela T. and Lähdesmäki H., Self Organization and Missing Values in SOM and GTM, Neurocomputing, Vol. 147, 2015, pp.60–70.

[20] Hazrati S.Y. and Datta B., Self Organizing Map Based Surrogate Models for Contaminant Source Identification Under Parameter Uncertainty**,** International Journal of GEOMATE, Vol. 13, Issue 36, Aug 2017, pp.11-18.

[21] Saitoh F., An Ensemble Model of Self-organizing Maps for Imputation of Missing Values, 2016 IEEE 9th International Workshop on Computational Intelligence and Applications, Hiroshima, Japan, 5 November 2016, pp. 9-14.

[22] Keerthika G., Feature Subset Evaluation and Classification using Naive Bayes Classifier, Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org Vol. 1, Issue 1, March (2015), pp. 22-27.