ASM Sc. J., 13, 2020
https://doi.org/10.32802/asmscj.2020.sm26(5.23)

Missing Data Imputation with Hybrid Feature Selection for Fertility Dataset

Mohamad Faiz Dzulkalnine, Roselina Sallehuddin*, Azlan Mohd Zain, Nor Haizan Mohd Radzi, Noorfa Hazlinna Mustaffa

School of Computing, Universiti Teknologi Malaysia, Skudai, 81300 Johor, Malaysia

Missing values poses a great concern in medical analysis as it may alter the result of analysed data and cloud the judgement of the medical practitioner which ultimately affecting the precise treatment a patient should receive. Even though there are many imputation methods that have been developed, the main issues with missing values such as accuracy and bias in prediction remain unsolved. In this paper, Fuzzy c-means (FCM) is employed as the imputation method. However, FCM does not consider the factor of irrelevant features. Noise and redundant data in the irrelevant features can reduce the accuracy of imputation and increase the computational time of FCM. An approach to tackle this problem is by using a feature selection method. By removing features that are irrelevant, the accuracy of imputation can be increased. Therefore, in this study, a hybrid imputation model Principal Component Analysis-Support Vector Machines-FCM (PCA-SVM-FCM) is proposed. The effectiveness of the proposed model is tested on a medical dataset which is Fertility dataset. Its performance is then validated by comparing it with SVM-FCM. Experimental result demonstrated that the proposed model performs better than SVM-FCM by producing a much lower error in estimation when tested using RMSE and MAE. The proposed model was then further verified by using Thiel's U test and producing low U value that indicates it is sufficient and significant. Therefore, PCA-SVM-FCM can be a feasible imputation tool to assist medical practitioner to obtain a reliable and better data analysis result.

Keywords: Missing data, feature selection, imputation, classification, backward selection

I. INTRODUCTION

Missing data occurs due to several factors including incorrect entering of data, non-responsive answers in surveys, data corruption, and many more. It poses several problems in data analysis and could affect the performance of machine learning and data mining. One of the main concerns of missing data in data analysis is that it could cause the loss of precision in estimation as there are fewer data to correlate and bias in the data distribution. Missing data is very critical in medical analysis as it could affect the process of creating a correct diagnosis from case records. This could mislead the judgement of the medical practitioner and ultimately a correct care a patient should be given. For instance, in fertility diagnosis, it is vital to analyse the patients IVF data and assessing the possible success rate of the treatment which could help the gynaecologist to suggest the appropriate fertility treatment for a couple to have a baby. The presence of missing data in the fertility record will contribute to lower amount of records available for data analysis. This will in turn give misleading result in

*Corresponding author's: roselina@utm.my

the analysis and ultimately giving incorrect treatment to the hopeful fertility patients

Conventionally, the missing data are usually handled by ignoring, deleting, or replace by mean estimation. However, these methods will reduce accuracy and introduce bias in thedata analysis. The bias in the data will also happen particularly when the missing data rate of is high. Simply deleting the missing data resulted bias which later generates inaccurate results (Panigrahi and Mishra, 2014).

Therefore, reliable methods that can predict the missing value systematically are urgently needed. The missing data should be predicted by using a proper imputation method. Researchers have developed several imputation methods in the past such as Singular Value Decomposition impute (SVD impute), Bayesian Principal Component Analysis (BPCA), and K Nearest Neighbour (KNN). However, the accuracy of these methods relies on the type of data whether it is homogenous or heterogeneous type of dataset.

In 2013, a hybrid imputation model which integrates FCM with support vector regression and genetic algorithm is proposed in (Aydilek et al., 2013). This method introduces a training phase so that the accuracy of imputation is improved by continuously reducing the error between the imputed dataset with the trained dataset. However, the author states that the performance of the proposed method could be improved if a feature selection method were implemented before the training phase. Feature selection will remove any irrelevant or redundant feature that could jeopardise the imputation performance of the imputation method (Urbanowicz et al., 2018).One of the most widely used feature selection methods is Principle component analysis (PCA) (Gu et al., 2018). PCA is well known for its ability to identify relevant features from the dataset. The advantage of removing irrelevant features is that it could reduce the computational time while improving the performance of the machine learning methods. Published study demonstrated that PCA is better than genetic algorithm, information gain ratio, and relieve attribute evaluation function (Koutanaei et al., 2014). The same study also developed a hybrid of PCA and SVM model which produce better classification accuracy compared to PCA with Decision Tree, Naïve Bayes, and ANN respectively. This shows that the hybridization of PCA and SVM able to yield a good classifier performance.

Thus, this work proposes PCA is used as the feature selection method before the data imputation phase. Here, SVM implemented as classifier to measure the accuracy of PCA's selected features in order to rank them before the next phase which is the imputation process. Therefore, the objective of this study is to propose a new imputation method by FCM with hybrid feature selection using PCA and SVM. The remaining of this paper is structured as follow: Section 2 outlines on the materials and methods used in this study. In section 3, the obtained results from the experiment are presented and discussed. Finally, conclusion and future works are presented section 4.

II. MATERIALS AND METHODS

A. Dataset

In this study, the proposed model is validated using public fertility dataset from the UCI machine learning data repository. The fertility dataset consists of 100 instances represented by 10 features where the predictive class is the diagnosis of the fertility, "0" represents normal and "1" for altered. There are 9 input features: Season in which the analysis was performed ($x1$), Age at the time of analysis ($x2$), Childish diseases ($x3$), Accident or serious trauma ($x4$), Surgical intervention ($x5$), High fevers in the last year ($x6$), Frequency of alcohol consumption ($x7$), Smoking habit ($x8$) and Number of hours spent sitting per day($x9$). The output is a predictive class (y) that has a value of "1" or "0". For the classification, the dataset is divided into two partitions: training and testing. There are three categories of data ratio in these partitions which are 50-50, 70-30, and 80-20 (Kannagara et al., 2018). The purpose of these percentages used is to make sure equal class distribution of the dataset. The different training testing percentages also to make sure that the factor of data imbalance will not affect the overall classification accuracy. The three different training testing partitions are chosen as it is important to have a much larger

portion of the dataset as the training set so that it will reduce the computational time during the testing phase and increasing the testing accuracy. If there is less information during the training phase, such as 20-80, 30-70, or 40-60, the model might become under fit and gives incorrect results when new data arrives in the testing phase. In the result section, the outcome for using different training-testing partitions is presented and discussed.

B. Performance Measurement

The performance of the proposed method is tested by using four different types of performance measurement methods. The proposed model was divided into two phase which is Feature Selection (FS) phaseand Imputation phase andeach phase have a different performance testing mechanism.

For the feature selection phase, classification accuracy is used as the performance criterion. Meanwhile in the imputation phase, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are employed to measure the rate of imputation error by the proposed hybrid imputation model, PCA-SVM-FCM. Finally, in order to validate the performance of the proposed model, Thiel's U test is applied.

C. Proposed method PCA-SVM-FCM

Figure 1 shows the framework of the proposed PCA-SVM-FCM model that combined PCA-SVM as feature selection and FCM as imputation model. In the FS phase, PCA works as ranking feature scheme that ranked each feature based on PC score.

In general, PCA works in the following steps. Firstly, the mean for each dimension of the dataset is obtained and subtracted from the dataset. Then, the covariance matrix is obtained by calculating its corresponding eigenvalues and eigenvectors. The eigenvalues which is also known as Principal Component (PC) scores correspond to the eigenvectors is the principal components of the new dimension of the dataset. After the eigenvectors are identified, the subsequent step is to rank them by eigenvalues from highest to lowest. The eigenvalues ranking will give the features order of relevancy.

Here, the general PCA steps are translated into following steps. First PCA will calculate the score for each feature and then descending rank all the input features according to PC scores. The feature with the highest PC scores which is the most significant features is ranked first while the feature with the lowest PC scores which is the least significant features is ranked last and deleted. Each time a feature is deleted, SVM classifies the performance of the reduced dataset and obtain its accuracy. The deletion of the least significant feature is repeated until there is no improvement in SVM's accuracy performance. The remains features are the significant features that can be used as input to represent the hidden patterns in fertility dataset. Then the best set of features selected by PCA-SVM acts as an input for Fuzzy c-means to predict the missing values in phase 2.To test the

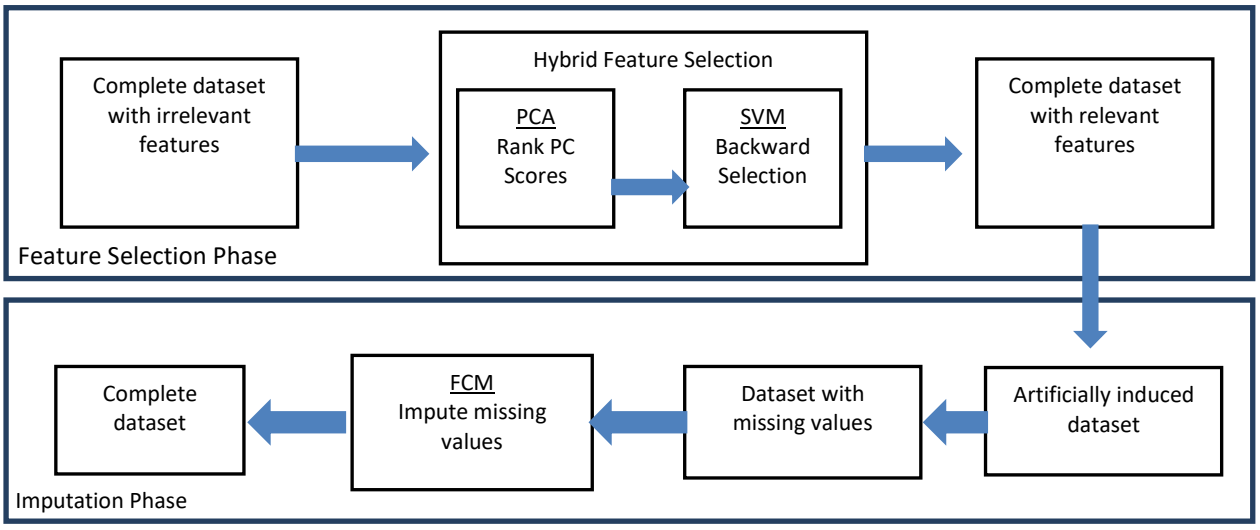

Figure 1. Proposed hybrid imputation model, PCA-SVM-FCM

robustness of the proposed model, the missing values are artificially induced in the datasets with increasing missing rate which are 1%, 5%, 10%, 15%, and 20% (Liu et al., 2016; Purwar and Singh 2015). Then, FCM is employed to impute the dataset. The process of estimating the missing values is repeated until it fulfils the condition set by FCM which is the maximum number of iteration or when the objective function of FCM is achieved.

FCM algorithm functions as follows. Initially, FCM will determine the cluster centre for each cluster. Every data object has a membership function to determine the degree of belonging of the data object to each cluster. The missing values are computed by calculating the data object degree of membership to the clusters and the value of centroids for each cluster it belongs to. The process will terminates when the maximum number of iterations set at 100 is reached, or when the improvement between two consecutive iterations is less than the minimum amount of improvement specified (0.0001).The study utilized Matlab version R2011a to perform the implementation and testing of the proposed model.

III. RESULTS AND DISCUSSIONS

In this section, discussion on the result is separated into three parts which is feature ranking by PCA followed by the SVM's classification result based on backward sequential selection. The final part will cover on the imputation performance of the proposed model.

A. Feature Ranking by PCA

First, the features in the dataset are ranked in descending order according to their respective PC scores. The higher the PC scores, signifies higher feature relevancy. Consequently, features will lower PC scores will be deemed irrelevant. Figure 2 shows the ranking for each feature with its respective PC scores.

x8 (0.0971)>**x7** (0.0635)>**x2** (0.0027)>**x6** (-0.0018)>**x3** (-0.0062)>**x5** (-0.0066)>**x9** (-0.0194)>**x4** (-0.0527)>**x1** (-1.6322)

Descending Order

Figure 2. Ranking PC Score by PCA

From the above ranking, the most important feature that affects the fertility is Smoking habit (*x*8) with PC scores of 0.0971 while the features with the lowest PC scores with -1.6322 that is the least significant input is Season in which the analysis was performed (*x*1). The reason why *x*1 ranked the lowest is that the features might be plagued with noise or redundant data compared to the other features.

B. Determination of Significant Input using SVM with Backward Selection

To obtain the optimal number of significant inputs, SVM with backward selection is implemented. These ranked inputs from PCA will be fed into SVM classifier and the SVM classification performance is recorded. To identify which features to be deleted, the performance criterion selected in this study for SVM is classification accuracy. Then the least significant input that is *x*1 is deleted and the remaining eight inputs are used in SVM classification and its performance is recorded.

This process is repeated until there is no change in SVM accuracy. Here, the SVM accuracy increased until seven inputs and started to decrease when six input is used. Therefore, the final optimum significant inputs are seven and they are *x*8, *x*7, *x*2, *x*6, *x*3, *x*5, and *x*9. Table 1 shows the classification performance of PCA-SVM and SVM. The classification accuracy of the original dataset by SVM are denoted by $\text{Fertility}_{}$ is to obtain the benchmark performance with the irrelevant features are still present while PCA-SVM denotes by Fertility_{PCA-S} is the classification result of the reduced dataset. The features are first trained by SVM by using 50%, 70% and 80% of the data and the rest of the data will be used for testing. The classification accuracy

Table 1. Classification accuracy for SVM and PCA-SVM for Fertility dataset at three different training testing partitions

Methods	Classification accuracy (%)					
	50-50	70-30	80-20	Average	Min	Max
Fertility_{SVM}	84	85.6	87	85.6	84	87
$\text{Fertility}_{PCA-SVM}$	92	90	89	90.3	89	92

Table 2. RMSE and MAE comparison for PCA-FCM and SVM-FCM at five different rates of missing values

	1%		5%		10%		15%		20%	
	PCA-FCM	**SVM-FCM**	**PCA-FCM**	**SVM-FCM**	**PCA-FCM**	**SVM-FCM**	**PCA-FCM**	**SVM-FCM**	**PCA-FCM**	**SVM-FCM**
RMSE	0.015	0.016	0.028	0.032	0.052	0.053	0.052	0.053	0.068	0.069
MAE	0.002	0.003	0.007	0.008	0.019	0.020	0.018	0.019	0.034	0.037

shown in the table proved that the performance of SVM increased significantly by implementing PCA as the feature selection method. As can be seen from Table 1, the best result is obtained from the 50-50 partition where PCA-SVM returns the highest classification accuracy which is 92%.

The reason PCA-SVM perform best at the 50-50 partition is that the fertility dataset is a very small dataset with only 100 attributes. Thus, at the 80-20 partition, there are too few data for the testing and could affect the generalization ability of SVM. PCA-SVM also performs better than SVM in all data partition as the average classification accuracy for all the training-testing partition is much higher than SVM. The better performance of PCA-SVM illustrated the importance of feature selection in classification. By removing irrelevant features that are plagued by noise or redundant data, SVM able to classify the dataset by much higher accuracy.

C. Imputation Phase

In the imputation phase, the set of relevant features that were obtained from the previous phase are artificially induced with missing values with five types of percentages which are 1%, 5 %, 10%, 15% and 20 %. This is to test the robustness of the imputation method with increasing rate of missing values. The performance of the FCM is evaluated by using three types of performance measurement methods such as RMSE, MAE and Thiel's U test. In order to achieve more comprehensive result, the experiments were repeated ten times and then its average result is calculated. The proposed imputation model was also compared to SVM-FCM in order to observe whether with the implementation of feature selection has an effect to its imputation accuracy. The first performance measurement to determine the imputation accuracy is RMSE. Table 2 showed the improved performances of the proposed model when compared to SVM-FCM.

PCA-FCM produced much lower RMSE value compared to SVM-FCM at each missing value rates. This is due to PCA ability to reduce the dataset dimension compared to the full dataset.

In addition to RMSE, this study also compared the performance of PCA-FCM with SVM-FCM by using MAE. In Table 2, it can be observed the obtained MAE result is also in line with the RMSE result in Table 1. This is crucial as both RMSE and MAE are error-based performance measure metric. From the table, we can conclude that the proposed model produced minimal error compared to SVM-FCM.

Next the Thiel's U test results are tabulated in Table 3. Thiel's U test is a relative accuracy measurement that compares the real data with the results of forecasting with minimal historical data. Thiel's U test is measured by evaluating the U value. U value closer to zero demonstrates better forecasting accuracy while U value closer to 1 indicates bad forecasting accuracy. In this study, PCA-FCM produced near zero U value that further validates the predictive ability of the proposed model.

Table 3. Thiel's U test result for the imputation performance of PCA-SVM-FCM

Thiel's U test	**1%**	**5%**	**10%**	**15%**	**20%**
	0.003	0.005	0.015	0.015	0.018

Therefore, the experimental results obtained from this supports the claim from the previous study (Aydeliket al., 2015; Sefidian and Daneshpour, 2019) that is to improve the capability of FCM imputation, the redundant and irrelevant features need to be identified and remove from the input list.

IV. SUMMARY

Missing values in medical dataset is a critical as it could reduce the analysis accuracy and introduce bias when deducting a conclusion from a case files. Loss of accuracy and bias in prediction are worsen due to the presence of irrelevant features in dataset. This further diminishes the efficiency of the imputation method as noise or redundant data in those features will affect the calculation of the missing values. Feature selection is known to be able to alleviate the issue by removing irrelevant features. In this study, a new hybrid imputation model, PCA-SVM-FCM has been proposed. Here, PCA is used to identify relevant features in the dataset and ranked them based on their priority. SVM is then used to classify the features to evaluate which features deemed irrelevant. The optimum number of features obtained from PCA-SVM model is seven and further decrease the features resulted lower classification accuracy. Later, these selected significant features are used for data imputation using FCM. The effectiveness of the proposed method was tested on one dataset which is Fertility dataset. The results obtained showed that PCA-SVM performs better compared to standard SVM by producing much higher classification accuracy.

The removal of irrelevant features contributed to the increase imputation performance of FCM in all tests. The proposed model is also validated by comparing its performance to SVM-FCM and proved that feature selection able to increase the performance of FCM when irrelevant features that affect the calculation of the missing values are removed. The promising predictive ability demonstrated by PCA-SVM-FCM that was supported by the result showed that it can be used in medical institutions to assist medical practitioner to obtain a better and more accurate diagnosis.

V. ACKNOWLEDGEMENT

This study is supported by GUP- tier 1, (Vot 16H57). Authors would like to thank Research Management Centre (RMC) Universiti Teknologi Malaysia, and ALI@S for the support in research activities.

VI. REFERENCES

Aydilek, I. B., &Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, *233*, 25-35.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: introduction and review. *Journal of biomedical informatics*.

Kannangara, M., Dua, R., Ahmadi, L., & Bensebaa, F. (2018). Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. *Waste management*, 74, 3-15.

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, *27*, 11-23.

Panigrahi, L., Das, K., & Mishra, D. (2014). Missing value imputation using hybrid higher order neural classifier. *Indian Journal of Science and Technology*, *7*(12), 2007-2014.

Purwar, A., & Singh, S. K. (2015). Hybrid prediction model with missing value imputation for medical data. *Expert Systems with Applications*, *42*(13), 5621-5631.

Sefidian, A. M., & Daneshpour, N. (2019). Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. *Expert Systems with Applications*, *115*, 68-94.

Gu, S., Cheng, R., & Jin, Y. (2018). Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft Computing*, *22*(3), 811-822.