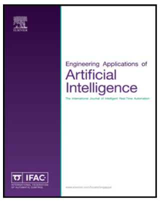
Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction[☆]

Ching-Hsue Cheng [*], Chia-Pang Chan, Yu-Jheng Sheu

Department of Information Management, National Yunlin University of Science and Technology, 123 University Road, Section 3, Douliu, Yunlin, 64002, Taiwan, ROC

ARTICLE INFO

Keywords:
Missing value
Imputation
Purity-based k nearest neighbor
Financial distress

ABSTRACT

Financial distress research often has missing values problems, and the different missing values handling techniques have an impact on the classification results. Furthermore, missing values handling in the data sciences is an important issue, and the different missing values handling approaches restrict on the application and performance of the classification. In missing values research, previous studies usually focused on the accuracy of classification, however, they address less the overall performance of the different missing degrees. To obtain better accuracy and maintain the integrity of data on the classification, this study proposes a purity-based k nearest neighbor algorithm to improve the performance of the missing value imputation. To verify, this study implemented different missing degree and different noise rate experiments for demonstrating the better performance because the proposed method is less affected by the noise. Furthermore, this paper also implemented MAR, MCAR, and MNAR type experiments, and compared the proposed method with the listed imputation techniques. Furthermore, this study practically collected Taiwan Economic Journal (TEJ) datasets as MNAR type missing values, and then employed the proposed purity-based k nearest neighbor algorithm to build a financial distress prediction model. Finally, this study compared the proposed imputation algorithm with common imputation methods and different classifiers, the results show that the proposed imputation algorithm obtains better accuracy and more stable in different missing degrees and noise.

1. Introduction

Financial distress (or financial crisis) is a business situation in which cash flow is not sufficient to pay a debt. Financial crisis prediction is an important and challenging research topic. Since 1966 (Beaver, 1966), many methods have been used to predict corporate bankruptcy and financial crisis, including artificial intelligence and statistical methods; many studies have shown that artificial intelligence is better than traditional statistical methods (Jerez et al., 2010). Additionally, the financial distress model predicts whether a company will fall into financial distress based on the recent financial data, which can be predicted by mathematical, statistical or data mining techniques (Sun et al., 2014). To allow enterprises, financial institutions and investors to take preventive or remedial actions as soon as possible before the financial crisis, it is necessary to build a method to warn of financial crisis.

In financial practice however, financial statement of enterprise is often quite limited, which makes modeling challenging and available data precious. In addition to the limited nature of data, the existing data are usually impaired by incomplete records. Therefore, the unavailability of these records particularly amply the problem of scarce data. Moreover, many standard statistical procedures require complete data. To build a good financial distress early warning model, this study proposes a novel missing value imputation method with a theoretical basis for stakeholders.

Handling missing values is performed during the pre-process of data mining. The pre-process is a necessary procedure for obtaining a better outcome. Without carefully handling missing values in pre-processing, the outcomes of analysis may distort the facts. Therefore, missing values handling is an important procedure in pre-processing. For handling missing values, many researchers have proposed various types of missing value handling techniques. However, most of the research focused on classification accuracy and ignored the effect of different missing degrees of data that could result in a questionable outcome. Simultaneously, the predicted values are also susceptible to outliers (or noise). Many missing values handling techniques removed outliers from the dataset, then either performed imputation or contained the outliers

[☆] No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.engappai.2019.03.003.

[*] Corresponding author.
E-mail address: chcheng@yuntech.edu.tw (C.-H. Cheng).

https://doi.org/10.1016/j.engappai.2019.03.003
Received 20 October 2017; Received in revised form 18 February 2019; Accepted 2 March 2019
Available online 13 March 2019

to perform prediction. One is a violation of the spirit of information science, and the other is the result of unreal output.

Addressing the financial distress data with missing values problem, many studies tend to use traditional statistical methods, such as the listwise approach (Allison, 2002), hot deck imputation (Andridge and Little, 2010) and cold deck imputation (Shao, 2000). Furthermore, financial distress focuses on two-class labels (health or distress). Based on the previously mentioned problems of outliers, artificial intelligence is better than traditional statistical methods. To obtain better results, many researchers proposed multiple imputation techniques, where distinct estimate techniques imputed missing values. However, removing instance types remains a concern, and they do not discuss the different degrees of missing values. Therefore, this paper proposes a new imputation algorithm to handle both different missing degrees and maintaining all instances. This paper has four contributions as follows:

(1) Propose a new imputation algorithm based on purity k nearest neighbors imputation (PkNNI) for missing values imputation, and demonstrate the effects of noise on the proposed imputation method.
(2) Compare the performance of different missing degrees for the parameter combination of the proposed imputation method.
(3) Compare the performance of different imputation methods with the proposed imputation method.
(4) Build a financial distress prediction model based on the proposed imputation techniques.

The rest of this paper is organized as follows: Section 2 describes the related work including financial distress, type of missing values, and imputation techniques. Section 3 introduces the concept and procedure of the proposed method. Section 4 is experimental framework, environment, datasets description, and experimental results. The conclusion is in Section 5.

2. Related work

In this section, the related literature and concept of missing values and k nearest neighbor technique are introduced in the following.

2.1. Financial distress

Financial distress (or called financial crisis) refers to the situation where a firm's cash flows are not enough to meet contractually required payment. Moreover, the financial distress model is to predict whether a company will fall into financial distress based on the recent financial data, can be predicted by mathematical, statistical or data mining techniques (Sun et al., 2014). There are two stages to predict financial distress. One uses the different financial features, and the other uses different classifiers used in building the prediction model (Lin et al., 2011).

Firstly, in financial features stage, financial ratios recognized as one of the most important factors affecting bankruptcy prediction, are used to develop prediction models (Beaver, 1966), a ratio denotes the mathematical relationship between one quantity and another such as current ratio, inventory turnover, EPS, and fixed asset turnover. Lin et al. (2011), they selected all 74 financial ratios, referred as the TEJ feature set, and combine this set with those 21 financial ratios recommended by previous research in business crisis prediction. Furthermore, the most usability of financial ratios was classified in seven broad categories: Liquidity, Asset Utilization, Long-Term Solvency, Profitability, Cash Flow, Market Valuation, and Size.

Secondly, in the classifier stage, various techniques have been proposed such as multiple discriminate analysis (MDA) and logistic regression in statistics, or neural network (NN), support vector machine (SVM) and decision tree (DT) in machine learning (Li et al., 2009). Li and Sun (2008) used a case-based reasoning (CBR) system with k nearest neighbor as a new measure model to forecast the financial distress of Chinese companies and compared it with two types of distance and inductive approaches. Ding et al. (2008) proposed a prediction model for forecasting financial distress based on support vector machines, and they found the best parameter value of the kernel function of C using grid-search techniques. Zhou et al. (2015) investigated the performance of different financial distress prediction models with feature selection, which was based on domain knowledge, data mining, and combined the domain knowledge and a genetic algorithm with a feature selection method that outperformed unique domain knowledge and unique data mining-based feature selection methods in AUC (area under the curve of ROC) performance.

2.2. Type of missing values

Applications of classification tasks often encounter data loss situations, called missingness. This can be caused for a variety reasons such as input errors, inaccurate measurement, equipment malfunction, measurement noise, data corruption, etc. This is known as unstructured missingness because the structures on the datasets do not have any implications. Missingness may also occur if attributes are not defined for all the data points. This is structured missingness or absence of features (Chechik et al., 2008). This paper focuses on the unstructured missingness problem. In unstructured missingness, Rubin (1976) developed a nomenclature for classifying the mechanism of missingness, and differed between three categories:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

The MAR type is when the missing values depend on the observed data points of an instance, and not on the unobserved data points, e.g., one participant is less likely to complete a depression survey, but this has nothing to do with their level of depression after accounting for maleness. If subjects who have missing values are a random subset of the complete sample of subjects, missing data are called MCAR, e.g., a participant flips a coin to decide whether to complete the depression survey. The MNAR refers to the case where missing values are subject to the unobserved attributes of an instance, e.g., participants with severe depression were more likely to be missing.

2.3. Imputation technique

In practice, there are two approaches for handling missing values, marginalization and imputation. Marginalization refers to the practice of excluding data instances with missing values. Simplicity is the main advantage of marginalization. However, the reduction of the statistical power and inability to perform comparison analysis (when pairwise deletion is used) are limitations. This leads to the loss of raw data and is ill-advised in applications where a sizable portion of the data has missing values and is therefore not within the scope of the study. Imputation aims to replace the missing values with predicted values for avoiding deletion of the instances or attributes and maintaining the integrity of the datasets. This paper focuses on the research of imputation techniques for missing values, and discusses the common missing imputation techniques.

Common imputation methods include zero imputation, average imputation, and class mean imputation (Donders et al., 2006). The adoption of common imputation methods is likely to reduce the variability of data. Additionally, mean imputation is affected by the presence of outliers, and therefore, median imputation is more appropriate in some cases and may create spikes in the data distribution. The detail about common imputation methods is described below.

(A) Zero imputation (ZI) fills the missing values with zero. The function of the zero imputation method is defined as Eq. (1):

$$ZI(I(i,j)) = \begin{cases} 0, & if\ I(i,j)\ is\ missing\ value \\ I(i,j), & otherwise \end{cases} \tag{1}$$

(B) Average imputation (AI) replaces the missing values with the averages of the corresponding attributes over the entire dataset. The expression of the average imputation is shown in Eq. (2):

$$AI(I(i,j)) = \begin{cases} \dfrac{\sum_{k=0}^{|S|} I(k,j)}{|S|}, & if\ I(i,j)\ is\ missing\ value \\ I(i,j), & otherwise \end{cases} \tag{2}$$

Where S∈X, X is an incomplete instance, and S is a complete instance. Function $I(i,j)$ represents the i*th* instance, and the j*th* attribute. If $I(i,j)$ is a missing value, the average from set S replaces the attribute missing value.

(C) Class mean imputation (CMI) is a slight modification of AI which replaces the missing values with the average of the attribute over all instances within the same class label as the instance being filled. The function of class mean imputation is defined as Eq. (3):

$$CMI(I(i,j),C) = \begin{cases} \dfrac{\sum_{k=0}^{|T|} I(k,j)}{|T|}, & if\ I(i,j)\ is\ missing\ value \\ I(i,j), & otherwise \end{cases} \tag{3}$$

where T∈X, X is an incomplete instance and T is a complete instance, and class belongs to C(i). The missing values are replaced by the average from set T, which is another complete instance and with the same class as label $C(i)$. $C(i)$ represents the class label for the i*th* instance.

Rubin (1987) proposed a multiple imputation (MI) procedure to replace each missing value based on a set of plausible values. Multiple imputation is a statistical method to analyze incomplete datasets, *i.e.*, datasets for which some attribute values are missing. The multiple imputation procedure has three phases (Yuan, 2011):

(1) The missing data are filled m times to generate m complete datasets.

(2) The m complete datasets are analyzed using standard procedures.

(3) The results from the m complete datasets are combined for the inference.

An advantage of the multiple imputation method is that it can be flexibly applied to many variables. It is valuable if multiple imputation can be adjusted to the conditioning set, or if the researcher can choose from different conditioning sets. However, the problems of multiple imputation are that the method fabricates data, which is not scientifically ethical, and it requires three phases which expend time and cost. Furthermore, multiple imputation, despite integrating over the imputed values, only corrects the standard errors because it assumes that there is no bias (Rubin, 1996, equation 2.6).

The kNN algorithm is also built on missing values imputation (Batista and Monard, 2003), namely, the k-nearest neighbor imputation (kNNI). In kNNI, the missing values in incomplete instances are replaced by the average of the corresponding attribute of its k nearest neighbors which instance was complete without missing values. The kNNI handles records with multiple missing values and takes into account the correlational structure of the data (Acuña and Rodriguez, 2004). However, this method tends to include noise and outliers as part of the predictive value, which leads to problematic predictive outcomes, thereby affecting the effectiveness of the classification. Troyanskaya et al. (2001) proposed a weighted k nearest neighbor imputation (WkNNI) as another imputation method based on the k nearest neighbor technique. In recent research, k nearest neighbors with mutual information (García-Laencina et al., 2009) is another technique used to solve the problem of missing values. This method uses kNNI with a feature-weighted distance metric based on mutual information techniques to improve the classification accuracy. Tsai and Chang (2016) proposed a two-step algorithm, which combined instance selection and missing values imputation, to filter out noisy data from the dataset to improve the final classifier performance. Amiri and Jensen (2016) proposed three missing values imputation methods based on fuzzy-rough sets and its extensions, including fuzzy-rough sets with nearest neighbor (FRNNI), vaguely quantified rough sets with nearest neighbor imputation (VQNNI) and ordered weighted average based rough sets with nearest neighbor imputation (OWANNI) to handle the missing values.

Datta et al. (2016) proposed a combination of both the k nearest neighbor technique and the penalized dissimilarity measure with feature weighting to handle the missing values problem, which can be directly applied to datasets with missing values, without any preprocessing. The missing values are adaptively imputed by using SOM and K-NN in classification according to context (Liu et al., 2016); they also used the ensemble classifier as credal classification. Moreover, a novel algorithm based on random forests with surrogate splits is proposed for handling missing values (Xia et al., 2017).

3. Proposed method

Because the missing values problem often uses the deletion approach in the financial distress field, it is possible to remove key information in datasets. Many studies use the traditional statistical methods in imputation techniques, such as the listwise approach (Allison, 2002), hot deck imputation (Andridge and Little, 2010) and cold deck imputation (Shao, 2000). Additionally, many missing values handling techniques remove outliers from the collected dataset, and artificial intelligence imputation is better than traditional statistical imputation (Jerez et al., 2010). Therefore, this study proposes an artificial intelligence imputation algorithm to improve the accuracy of different missing degrees and outliers in the dataset. This study employs the characteristics of average imputation and class mean imputation and proposes the KNNI method to estimate the missing values. Furthermore, this paper utilizes different noise rates to illustrate that better performance because the proposed method is less affected by the noise. In financial distress prediction, many researchers focus on the two-class label (health or distress). This study further expanded to a three-class label (health, alert and distress) to build a financial distress prediction model for datasets with missing values.

This paper proposes a novel KNNI imputation method. The proposed method can be divided into two parts, one is purity training, and the other is purity imputation:

(1) Purity training

The objective in the first part is to calculate the purity of each complete instance to obtain the purity of the i*th* instance, as $P\_training(i)$ and is defined as:

$$P\_training(i) = \sum_{s=0}^{k_1} vote(C(i), N_s) \tag{4}$$

where the C(i) is the ith complete instance from datasets X. k_1 represents the rigidity of purity, i.e., k_1 is used to determine the numbers of the nearest neighbors for purity calculations. N_s represents the s*th* nearest neighbor instance, and the function vote() is used to return the class label between instance C(i) and N_s to identify whether they are the same or not, which is expressed as:

$$vote(C(i), C(j)) \begin{cases} 1, & if\ Class(i) = Class(j) \\ -1, & if\ Class(i) \neq Class(j) \end{cases} \tag{5}$$

where the $Class(i)$ represents the i*th* class label from datasets X, to deduce whether the instance $C(i)$ is pure or not through comparing the class label $Class(i)$ and $Class(j)$.

For example, if instance Z has a positive $P\_training(i)$, the instance Z and its neighbors have homogeneous heights. In contrast, if instance Z has a zero or negative $P\_training(i)$, the instance Z is filled with other types of class labels, which means instance Z may be noise or an outlier. If instance Z is used as an estimate, it will affect the results of imputation.

(2) Purity imputation

After the first part, the missing values can be predicted through the complete instances and purity values. In the purity imputation, the proposed method enhances the traditional k nearest neighbor algorithm with the purity value, which is obtained from the first part. The imputation function is defined as follows:

$$M(i,j) = P\_imputation(i,j,k_2) = \frac{\sum_{S=1}^{k_2} Ref(S,j)}{k2} \tag{6}$$

where $M(i,j)$ represents the i*th* instance and its j*th* attribute, which is the missing value. k_2 represents the number of reference targets, *i.e.*, k_2 determines the numbers of nearest neighbors that are utilized to estimate the missing value, and $Ref(S,j)$ represents the s*th* nearest instance and its j*th* attribute, which is a collection that contains all positive purity instance information from the complete instance.

The proposed imputation method is different from the original k nearest neighbor imputation method. This study uses all complete instances for each attribute (including noise and outliers) to estimate the missing values, and the proposed imputation method separates the low purity instance values to avoid the effects of noise and outliers. i.e., this paper only used all positive purity instance information to estimate the missing values. When a purity value of the nearest complete instance was greater than zero, the instance joined the Ref set. Otherwise, if the conditions of the purity value were false, that instance would not be added. Filtering the low purity values of the instance from the sorted Ref set can avoid the effects of noise and outliers on prediction values, resulting in more stable imputation values as results.

Proposed procedure

To understand the proposed imputation method, the procedure of the proposed imputation method, which includes four steps, is shown in Fig. 1: data collection, data preprocessing, data imputation (contains two algorithms), and evaluation. The detailed steps are introduced step-by-step in the following.

Step 1 Data Collection

This step identified two types of datasets. One was the UCI dataset repository, and the other was the TEJ financial distress datasets. In the UCI datasets, missing values were artificially implanted, in different degrees and attributes. The main reason for not using data with missing values was the requirement to have total control over the missing data in the dataset. The datasets with no missing values were chosen for this study to simulate the missing situation in different missing degrees. In the TEJ financial distress dataset, the collected dataset contained missing values. The missing values were imputed and compared with other popular methods.

Step 2 Data preprocessing

This step used two different source datasets for data preprocessing. In the UCI datasets, due to the datasets having no missing values, we conducted the experiment by randomly removing some values from the original datasets. The missing values were simulated in all variables except the class attribute. The missing degree of the missing values was defined as follows:

$$Missing\ degree = \frac{number\ of\ missing\ attribute\ values}{number\ of\ (instances \times dimensions)} \tag{7}$$

Where instances denote the number of data point, and dimensions represent the number of attribute (included class). In the UCI dataset, 5%, 10%, 15%, 20%, and 25% as the missing degree were chosen to verify the performance of the proposed imputation method and compare with the listed imputation techniques. The process of generating missing values must maintain at least one value in each instance. If it is not, all of the imputation techniques will not be utilized to impute the missing values. Based on the probability of an attribute's data being missing is independent of both the attribute and the other observed attributes, according to Garciarena and Santana (2017), this study generates MCAR missing values as Algorithm 1. The MAR type is when the probability of data being missing is dependent on other attribute, based on Garciarena and Santana (2017), the algorithm of generated MAR missing values are modified as Algorithm 2.

Step 3 Data Imputation

In the data imputation step, the missing value was replaced by a new imputed value by using the proposed imputation techniques and other listed imputation methods. This study chose the most common imputation techniques including MI, ZI, AI, CAI and kNNI to compare with the proposed imputation method. The proposed purity k nearest neighbors imputation (PkNNI) was divided into two types, one used the class label to estimate the missing values, denoted as Type 1 PkNNI. The other did not use the class label, denoted as Type 2 PkNNI. The difference between Type 1 PkNNI and Type 2 PkNNI for the estimated results was observed. The detailed sub-steps of the proposed imputation method are described in the following.

Step 3-1. Purity computation

Firstly, to avoid different ranges of attributes affecting the distance calculation, all instances of an attribute were normalized in the same scale based on the attribute meaning, such as benefit (max goal) and cost (min goal) scaled to [0, 1], and Z-score scaled to [−1, 1]. Secondly, based on each data point must at least has one attribute value existed, each imputation of attribute (or class), the dataset was split into incomplete instances with missing values and complete instances without missing values. For example, if the dataset has 11 attributes, after generated MAR and MCAR type of missing values, for each attribute, we partition once (one time) the generated missing dataset into incomplete instances and complete instances. Hence, this step repeated 11 times partitions, then, the purity of each complete instance was calculated. The instance x found its k*th* nearest neighbors in complete instances via the Euclidean distance, and the k is the parameter of k*th* purity (symbol as k1). Once the k nearest neighbors were determined, the purity of instance x was obtained by aggregating the votes of its neighbors. If instance x and its k*th* nearest neighbors had the same class label, the purity increased by one, else, the purity decreased by one. Through the purity computation, the instance x determined whether the instance was an imputation basis. The purity computation algorithm is shown in Algorithm 3.

We use Fig. 2 as an example to illustrate step 3-1. Let the parameter k1 be 5, to obtain the purity value of instance i (blue triangle). This step calculated its nearest neighbors x1, x2, x3, x4 and x5, then compared the class label. Instance i belonged to the type blue class, so its ordering of nearest neighbor was x1 (yellow circle), x2 (blue square), x3 (blue square), x4 (yellow circle) and x5 (blue square). According to the proposed method, the purity of instance i added the cardinality of the same class label for the neighbor and subtracted the cardinality of the different class label for the neighbor. That is, homogeneous nearest neighbors increased one, else, heterogeneous nearest neighbor subtracted one. Hence, the purity of this instance i was $3 - 2 = 1$.

After the purity computation, the instance x determines whether the instance is an imputation basis for the next step, 3.2.

Step 3-2. Imputation

With complete instances and its purity list, we replaced the missing values using the k nearest neighbor instances with a positive purity. An instance with a positive purity value indicated that the instance was highly homogeneous with its neighbors and could be used as a reference instance, else, a negative purity value indicated that an

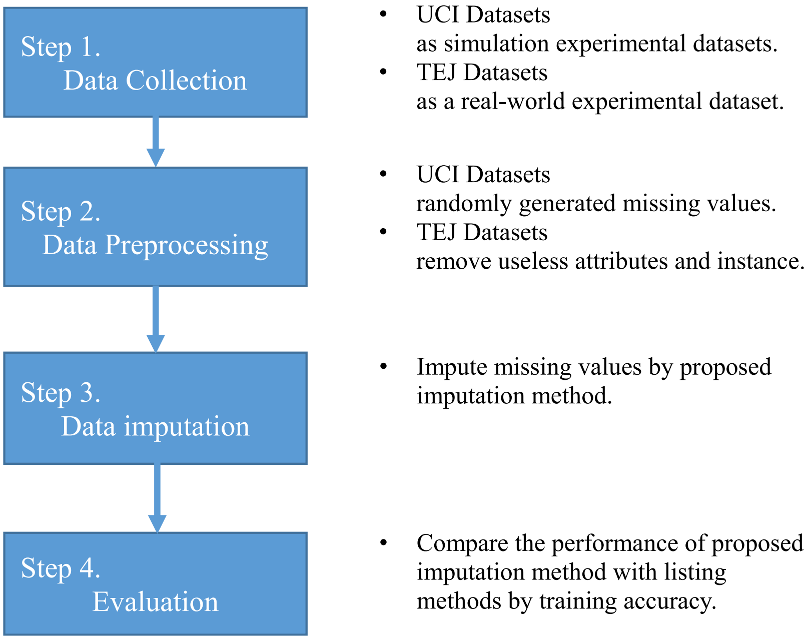
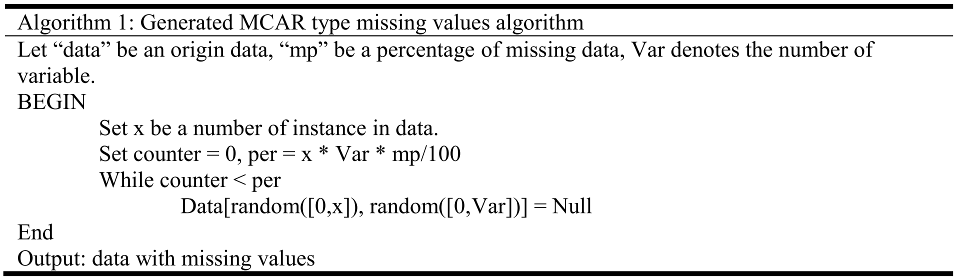
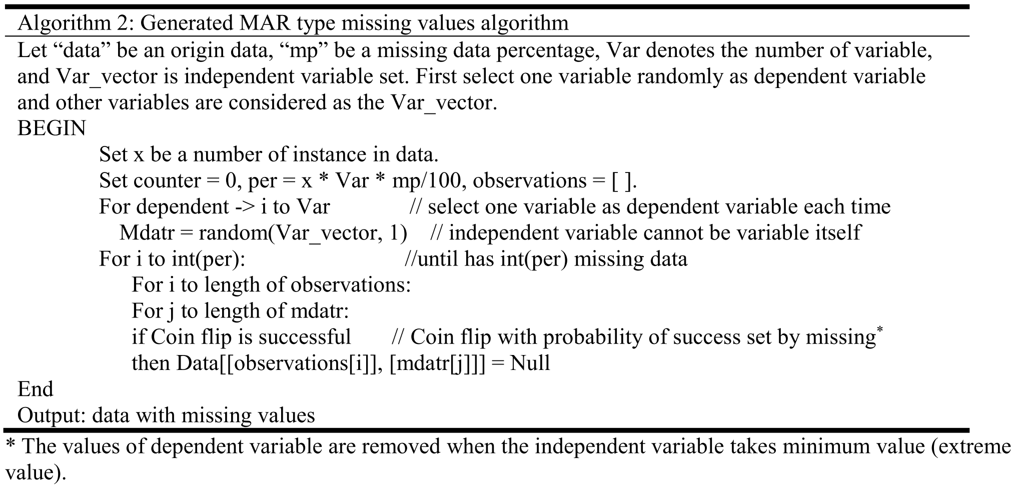

Step 1.
Data Collection
Step 2.
Data Preprocessing
Step 3.
Data imputation
Step 4.
Evaluation
• UCI Datasets
as simulation experimental datasets.
• TEJ Datasets
as a real-world experimental dataset.
• UCI Datasets
randomly generated missing values.
• TEJ Datasets
remove useless attributes and instance.
• Impute missing values by proposed imputation method.
• Compare the performance of proposed imputation method with listing methods by training accuracy.

Fig. 1. The procedure of proposed imputation method.



```
Algorithm 1: Generated MCAR type missing values algorithm
Let "data" be an origin data, "mp" be a percentage of missing data, Var denotes the number of
variable.
BEGIN
        Set x be a number of instance in data.
        Set counter = 0, per = x * Var * mp/100
        While counter < per
                Data[random([0,x]), random([0,Var])] = Null
End
Output: data with missing values
```



```
Algorithm 2: Generated MAR type missing values algorithm
Let "data" be an origin data, "mp" be a missing data percentage, Var denotes the number of variable,
and Var_vector is independent variable set. First select one variable randomly as dependent variable
and other variables are considered as the Var_vector.
BEGIN
        Set x be a number of instance in data.
        Set counter = 0, per = x * Var * mp/100, observations = [ ].
        For dependent -> i to Var            // select one variable as dependent variable each time
          Mdatr = random(Var_vector, 1)    // independent variable cannot be variable itself
        For i to int(per):                    //until has int(per) missing data
            For i to length of observations:
            For j to length of mdatr:
            if Coin flip is successful       // Coin flip with probability of success set by missing*
            then Data[[observations[i]], [mdatr[j]]] = Null
End
Output: data with missing values
```
* The values of dependent variable are removed when the independent variable takes minimum value (extreme value).

instance had a high degree of uncertainty with its neighbors so it may be an outlier or noise. Fig. 3 illustrates:

Set the parameter $k_2 = 3$. To impute the instance j, which contains missing values, it calculated its nearest neighbors, which had a positive purity value.

In Fig. 3(a), considering the different class labels, the inner digit of the triangle and circle denote impurity values, and the inner X of the triangle represents a missing value. We found that one nearest neighbor had a purity value of −1. The −1 indicates this nearest neighbor should not be considered and to search for other nearest neighbors with a positive purity value and the same class label. Finally, the three instances with the positive purity values 1, 3, 2, were considered the nearest neighbors.

In Fig. 3(b), ignoring the class labels, the difference from Fig. 3(a) was that it considered the class label of the complete instance, as the purity values 1, 3, 5 were considered the nearest neighbors.

After searching the nearest neighbors of the positive purity values, the proposed KNNI imputation method has two type of imputation

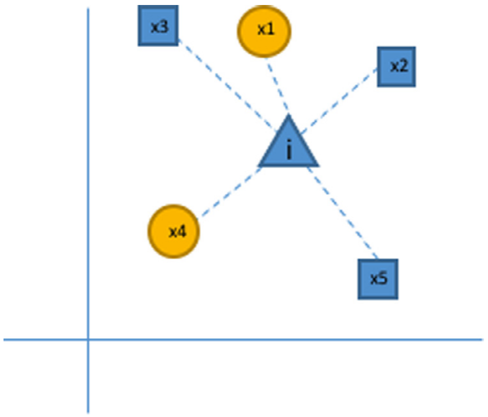


Fig. 2. Example of proposed purity computation.

computation. Type 1 is class mean imputation, it based on the same class label of all complete instances to compute class mean as imputation value as Fig. 3(a). Type 2 utilizes all the complete instances to compute the average imputation, it ignores the class label as Fig. 3(b). The imputation algorithm is listed in Algorithm 4.

Step 4 Evaluation

In the evaluation step, the performance of the classifier was calculated using the confusion matrix (Sammut and Webb, 2011). Accuracy is a popular metric. It refers to the ability of the model to predict the class label correctly, and it is the proportion. It is defined as follows:

$$accuracy = \frac{\sum_{i=1}^{n} N_{ii}}{\sum_{n=1}^{n} \sum_{j=1}^{n} N_{ij}} \tag{8}$$

where N_{ij} represents the number of instances belonging to class $A_{i\_}$ but are classified as class $A_{\_j}$, and N_{ii} denotes the number of instances, which were classified to the correct class. This study used training accuracy as the evaluation indicator because it shows the integrity of the data after being imputed.

4. Experiments and results

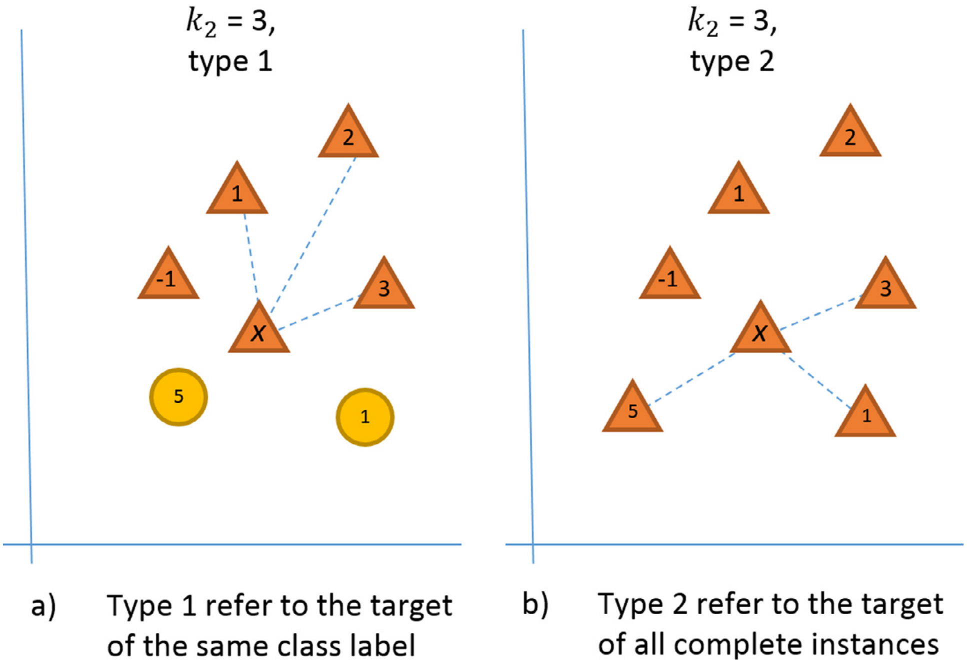
To verify the effectiveness of the proposed method, this study implemented the noise experiment to demonstrate that the better performance due to the proposed method is less affected by the noise. The experiment procedure followed Section 3's proposed procedure to demonstrate the effects of the proposed method for the UCI datasets in the MAR and MCAR experiments, and the practically collected financial dataset in the MNAR experiments. Eight different types of datasets were chosen from the UCI repository to verify the proposed imputation method for MAR and MCAR, and the practically collected TEJ datasets were employed for further verification and comparison for MNAR. Because the practical financial statement data is a very important decision-making tool for investors and governors, the impact of the financial statement data for investors is similar to medical data for a patient. Therefore, the authors set the missing value of the practical financial statement to the MNAR type. In the following, this paper introduces the experimental environment, noise in MAR and MCAR experiments for the UCI datasets, and the TEJ experimental datasets for MNAR.

4.1. Experimental environment

The experiment was implemented in Python (Python 2.7 version) on an Intel i7-3770k, 3.5 GHz CPU, with Ubuntu 14.04 LTS Operating system. This study used k = {3, 5, 7, 9} for kNNI, and the parameter of the proposed method PkNNI, set k1 = {3, 5, 7, 9}, k2 = {3, 5, 7, 9}. To compare the feasibility of the proposed imputation method, this study applied five common and well-known classifiers to compute the classification accuracy, and calculated the average training accuracy as a comparison indicator for verifying the feasibility of the proposed imputation method. The different imputation methods used in the experiment with the configuration parameters are listed in Table 1, and the five classifiers are listed in Table 2.

4.2. MAR and MCAR experiments

First, this experiment utilized different missing degrees of data (5%, 10%, 15%, 20%, and 25%) to run eight imputation methods



a) Type 1 refer to the target of the same class label

b) Type 2 refer to the target of all complete instances

Fig. 3. The impurity imputation example.

```
Algorithm 3 :Puriy_training
Let C = {x_1, x_2, …, x_n} be a set on n complete instances.
BEGIN
        Set $k_1$, 1 ≤ $k_1$ ≤ n.
        Initialize i = 1.
        DO WHILE (i ≤ n):
                Initialize j = 1.
                DO WHILE (k-nearest neighbors not found):
                        Compute distance from $x_i$ to $x_j$.
                        IF (j ≤ $k_1$) :
                                THEN Include $x_j$ in the set of k-nearest neighbors.
                        ELSE IF ($x_j$ is closer to $x_i$ than any previous nearest neighbor):
                                THEN Delete farthest neighbors in the set,
                                Include $x_j$ in the set of k-nearest neighbors.
                        END IF
                        Increment j.
                END DO WHILE.
                Determine the main class represented in the set of k-nearest neighbors.
                Compute the purity value of $x_i$.
                Increment i.
        END DO WHILE
END
Output: purity value of all complete instances
```

```
Algorithm 4: Purity_imputation
Let C = {$x_1$, $x_2$, …, $x_n$} be a set on n complete instances, M = {$y_1$, $y_2$, … , $y_m$} be a set on m incomplete
instances.
BEGIN
        Set $k_2$, 1 ≤ $k_2$ ≤ n.
        Initialize i = 1.
        DO WHILE (i ≤ m):
                Initialize j = 1.
                DO WHILE (k-nearest neighbors not found):
                        Compute distance from $y_i$ to $x_j$, by the observe attributes.
                        IF (Purity of $x_j$ is positive):
                                IF (j ≤ $k_2$ and purity of $x_j$ is positive) :
                                        THEN Include $x_j$ in the set of k-nearest neighbors.
                                ELSE IF ($x_j$ is closer to $y_i$ than any previous nearest neighbor):
                                        THEN Delete farthest neighbors in the set,
                                        Include $x_j$ in the set of k-nearest neighbors.
                                END IF
                        END IF
                        Increment j.
                END DO WHILE.
                Replace the missing values in $y_i$ as the average of nearest neighbors
                Increment i.
        END DO WHILE
END

Output: instance without missing values
```

Table 1
The different imputation methods and parameter settings.

Imputation method	Parameter	Reference work
ZI	None	Donders et al. (2006)
AI	None	Donders et al. (2006)
CMI	None	Donders et al. (2006)
MI	None	Rubin (1987)
kNNI t1	k = {3, 5, 7, 9}	Batista and Monard (2003)
kNNI t2	k = {3, 5, 7, 9}	Revised from Batista and Monard (2003)
PkNNI t1	k1 = {3, 5, 7, 9}, k2 = {3, 5, 7, 9}	Proposed
PkNNI t2	k1 = {3, 5, 7, 9}, k2 = {3, 5, 7, 9}	Proposed

Note: t1 considers class label, t2 is no considering class label.

(as Table 1) and five classifiers (see Table 2) for MAR and MCAR. Second, the study uses $x_{ij}\pm2\sigma_j$ (the ith instance, jth attribute) to generate different noise rates (3%, 6%, 9%, and 12%) and combine different missing degrees (5%, 10%, and 15%) for calculating the estimated value of the eight imputation methods, and then compute the average accuracy using five classifiers for eight different UCI datasets in MAR and MCAR experiments. The generated noise of different ratios is described as Algorithm 5.

In this experiment, eight different types of UCI datasets were chosen to verify the performance of the proposed imputation method for noise

Table 2
The parameters of the five classifiers.

Classifier	Parameters	Reference
C4.5	Confidence factor: 0.25	Quinlan (1993)
Naïve Bayes	None	John and Langley (1995)
MLP	Hidden layers: (attributes + classes)/2 Learning rate: 0.3 Momentum rate:0.2 Validation threshold: 20	Mitra and Pal (1995)
Bayes network	None	Pearl (1998)
Random forests	Iterations: 100	Breiman (2001)

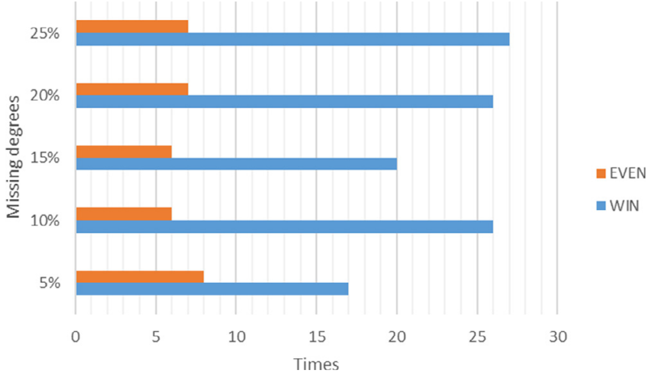


Fig. 4. The total wins and even of the proposed imputation for different missing degrees in MAR type. Note: "win" denotes the number of win for the best accuracy in 40 experiments (eight datasets, five classifiers).

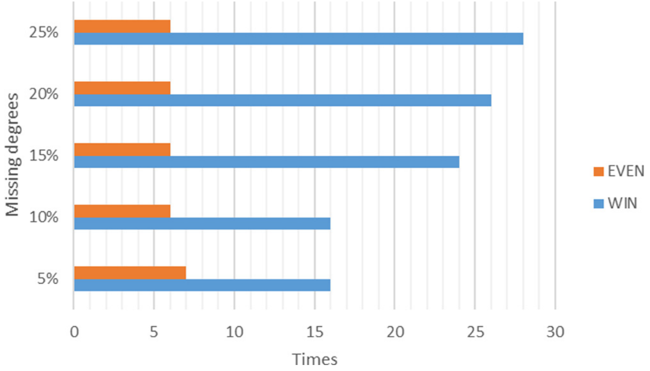


Fig. 5. The total wins and even of the proposed imputation for different missing degrees in MCAR type. Note: "win" denotes the number of win for the best accuracy in 40 experiments (eight datasets, five classifiers).

in MAR and MCAR. The eight UCI repository datasets use numeric attributes data, which do not contain missing values. The number of classes, number of attributes, and number of samples are described in Table 3.

The process of generating missing values must maintain at least one value in each instance. If it is not, all of the imputation techniques will not be utilized to impute the missing values. The experiment was conducted by removing values from the original datasets in MAR and MCAR. The missing values simulated all variables except the class attribute.

4.2.1. Missing degree experiment

This section utilized different missing degrees (5%, 10%, 15%, 20%, and 25%) to run eight imputation methods (as Table 1) and five classifiers (see Table 2) for MAR and MCAR. After the experiment, the proposed imputation method had the better accuracy in 40 experiments (eight datasets, five classifiers) for different missing degrees in MAR and MCAR as Figs. 4 and 5, respectively. We can see that the win ratio is greater than or equal to 1/2 ((win + even)/total) in all missing degrees, especially in cases of high missing degrees. Furthermore, the best accuracy of the proposed imputation methods in different datasets and datasets characteristic are shown in Table 4. In each dataset, the total comparisons are 25 experiments, combining five missing degrees and 5 classifiers (5 × 5 = 25). From Table 4, the proposed imputation method obtained the best win times in MAR and MCAR, except in the Ionosphere dataset.

Figs. 6–9 show that the proposed imputation had a better average accuracy in almost all of the five classifiers for eight different missing degrees. Fig. 6 shows that the proposed method obtained the better accuracy and had better results for different MAR missing degrees in the Banknote datasets. Fig. 7 shows that the proposed method had the best accuracy and maintained high stability for different MAR missing degrees in the Vertebral Column 3C datasets, but MI method has a larger fluctuation. It also shows that the proposed imputation method provided better accuracy in multiple class datasets. Figs. 8 and 9 show that the proposed method had the better results, and MI method has a larger fluctuation for different MCAR missing degrees in the Banknote datasets and Vertebral Column 3C datasets. It also shows that the proposed imputation method provided better accuracy in multiple class datasets.

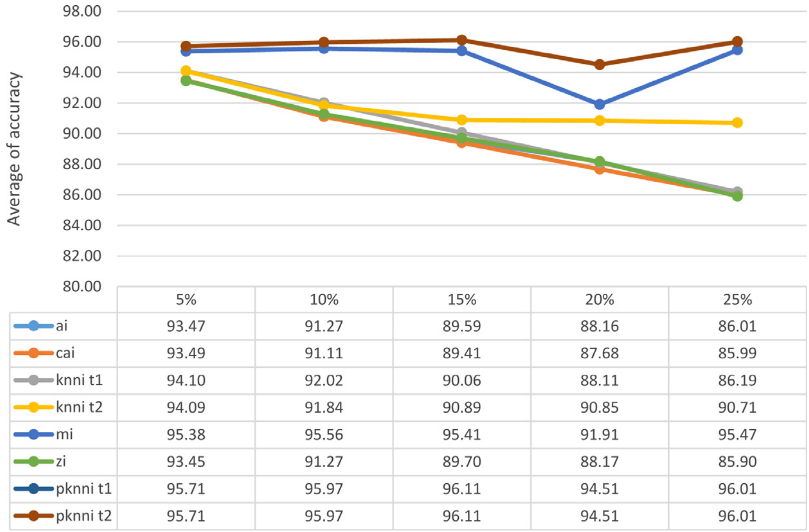
4.2.2. Experiment of mixed noise and missing degree

For verifying, the better performance because the proposed method is less affected by the noise. This sub-section uses statistical method, the Gaussian distributed noise is utilized to calculate standard deviation of each attribute, i.e., the experiments used $x_{ij} \pm 2\sigma_j$ (the ith instance, jth attribute) to generate noise of different ratios and different missing degrees of MAR and MCAR in each UCI open dataset.

The experiment was based on $x_{ij} \pm 2\sigma_j$ generate four noise rates (3%, 6%, 9%, and 12%), and produced three missing degrees (5%, 10%, and 15%). Then, eight imputation methods were employed to calculate the estimated value and compute the average accuracy of the five classifiers for the MAR experiments of eight different UCI datasets. After numerous experiments, due to too many experimental results, the detailed results are listed in Tables A and B of the Appendix. From Tables A and B of the Appendix, the 12 comparisons (four noise rate and three missing degrees, $4 \times 3 = 12$) for each dataset in MAR and MCAR are summarized in Table 5. Table 5 shows that the proposed imputation was better than other imputation methods in the MAR and MCAR experiments, except in the MI in the Ionosphere dataset. As shown in Tables A and B of the Appendix, all accuracies of PKNNIt1 were the same with PKNNIt2 in all mixed experiments. Therefore, the better performance due to the proposed method is less affected by the noise.

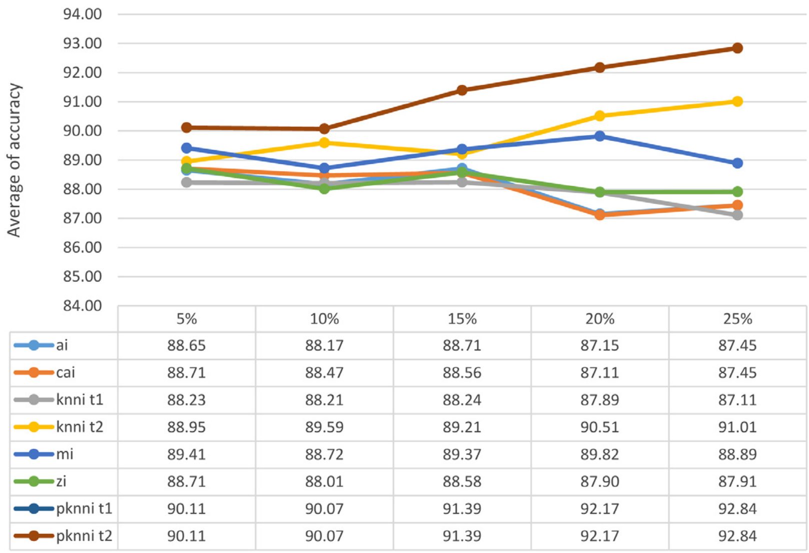
For a visual view, some figures of mixed noise and missing degree for MAR and MCAR experiments are shown in Figs. 10–13. From Tables A and B of the Appendix and Figs. 10–13, we can see: (1) when the noise rate rose, the accuracy declined, (2) when missing degrees increased, the accuracy decreased, and (3) in a lower noise rate and missing degrees, all imputation methods were robust. Fig. 10 shows the trends of mixed noise and missing degrees for the Banknote dataset in the MAR experiment, which shows the accuracy of all imputation

```
Algorithm 5: Generated noise data algorithm
Let D = be a dataset, N = number of instance in data, M = number of attribute in data, noise_per be a
noise percentage, noise_limit = (M * N) * noise_per, Dn = nth instance of dataset
BEGIN
        Set Noise_set = [],
                For i to M
                        Si = Standard deviation of ith attribute in data
                        Do WHILE count of Noise_set < noise_limit
                                Dn = random(0, N)
                                IF (Dn is not missing value and not in Noise_set)
                                        Dn = Dn +- (2 * SDi)
                                        Add Dn into Noise_Set
                                End IF
                        End do WHILE
                End For
End
Output: data with noise
```



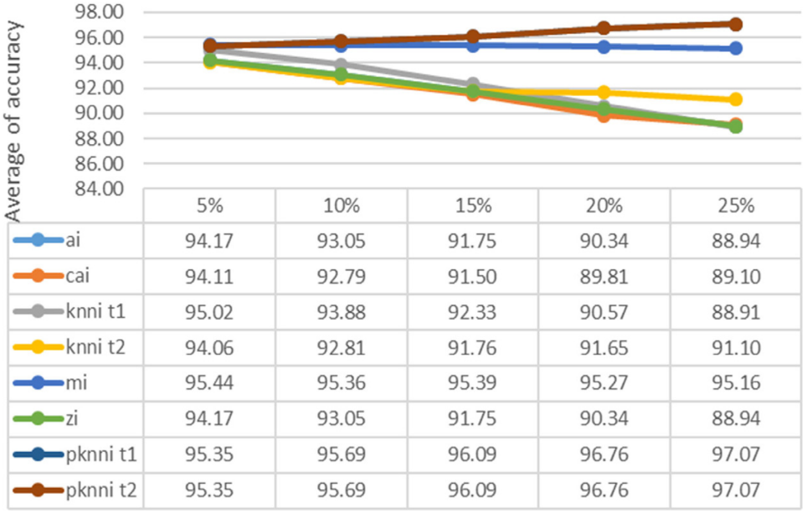
	5%	10%	15%	20%	25%
ai	93.47	91.27	89.59	88.16	86.01
cai	93.49	91.11	89.41	87.68	85.99
knni t1	94.10	92.02	90.06	88.11	86.19
knni t2	94.09	91.84	90.89	90.85	90.71
mi	95.38	95.56	95.41	91.91	95.47
zi	93.45	91.27	89.70	88.17	85.90
pknni t1	95.71	95.97	96.11	94.51	96.01
pknni t2	95.71	95.97	96.11	94.51	96.01

Fig. 6. The comparison results of different missing degrees of data in MAR under the Banknote dataset. Note: The results indicate that the proposed method has better accuracy and stability.



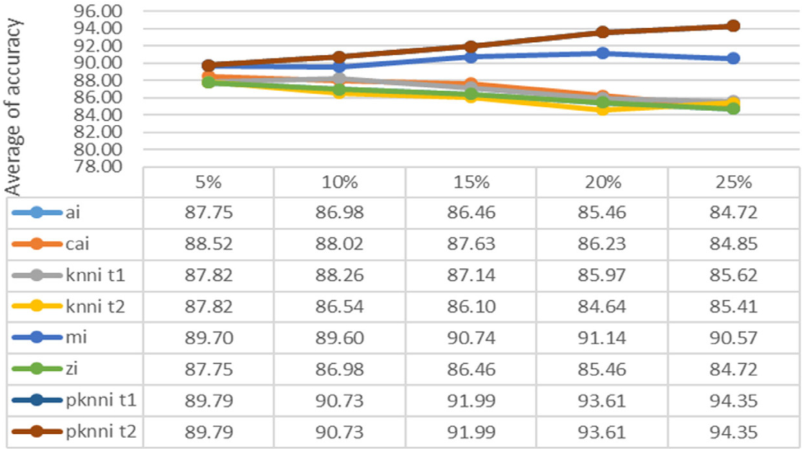
	5%	10%	15%	20%	25%
ai	88.65	88.17	88.71	87.15	87.45
cai	88.71	88.47	88.56	87.11	87.45
knni t1	88.23	88.21	88.24	87.89	87.11
knni t2	88.95	89.59	89.21	90.51	91.01
mi	89.41	88.72	89.37	89.82	88.89
zi	88.71	88.01	88.58	87.90	87.91
pknni t1	90.11	90.07	91.39	92.17	92.84
pknni t2	90.11	90.07	91.39	92.17	92.84

Fig. 7. The comparison results of different MAR missing degrees in the Vertebral Column 3C datasets. Note: The results indicate that the proposed method has better accuracy, but MI method has a larger fluctuation.



	5%	10%	15%	20%	25%
ai	94.17	93.05	91.75	90.34	88.94
cai	94.11	92.79	91.50	89.81	89.10
knni t1	95.02	93.88	92.33	90.57	88.91
knni t2	94.06	92.81	91.76	91.65	91.10
mi	95.44	95.36	95.39	95.27	95.16
zi	94.17	93.05	91.75	90.34	88.94
pknni t1	95.35	95.69	96.09	96.76	97.07
pknni t2	95.35	95.69	96.09	96.76	97.07

Fig. 8. The comparison results of different missing degrees of data in MCAR under the Banknote dataset. Note: The results indicate that the proposed method has better accuracy and stability.



	5%	10%	15%	20%	25%
ai	87.75	86.98	86.46	85.46	84.72
cai	88.52	88.02	87.63	86.23	84.85
knni t1	87.82	88.26	87.14	85.97	85.62
knni t2	87.82	86.54	86.10	84.64	85.41
mi	89.70	89.60	90.74	91.14	90.57
zi	87.75	86.98	86.46	85.46	84.72
pknni t1	89.79	90.73	91.99	93.61	94.35
pknni t2	89.79	90.73	91.99	93.61	94.35

Fig. 9. The comparison results of different missing degrees in MCAR under the Vertebral Column 3C datasets. Note: The results indicate that the proposed method has better accuracy, but MI method has a larger fluctuation.

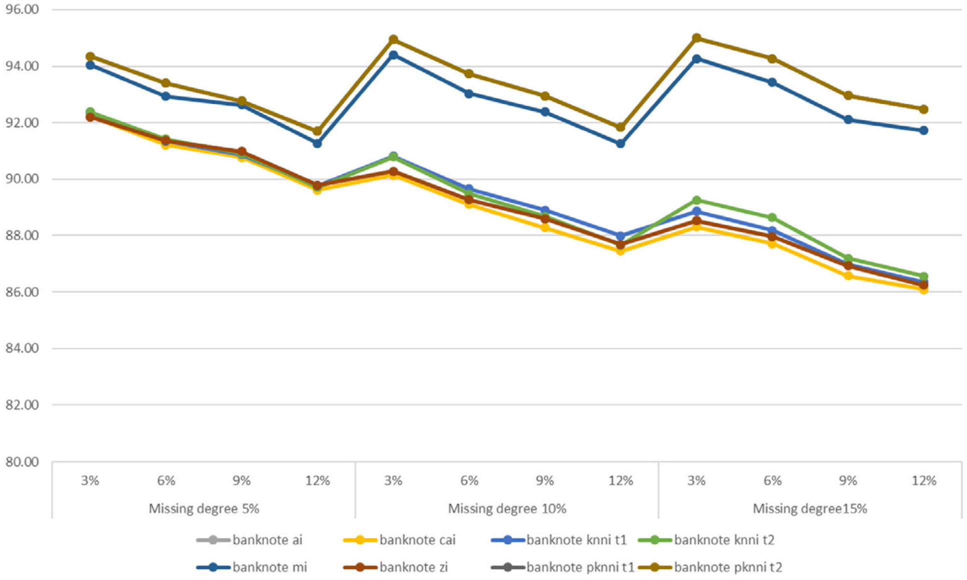


Fig. 10. The result of mixed noise and missing degree for the Banknote dataset in MAR. Note: The accuracy of all imputation methods quickly reduced when the noise rate increased.

Table 3
UCI datasets characteristic.

Name of dataset	Number of classes	Number of attributes	Number of samples
Banknote	2	5	1372
Blood Transfusion	2	4	748
Climate Model	2	19	540
Haberman	2	3	306
Ionosphere	2	34	351
Pima Indians Diabetes	2	8	768
Vertebral Column_2C	2	6	310
Vertebral Column_3C	3	6	310

Table 4
The number of wins for the proposed imputation for eight datasets.

Dataset	#Class	#Attributes	#Samples	#Win (MAR)	#Win (MCAR)
Banknote	2	5	1372	17	13
Blood Transfusion	2	4	748	14	14
Climate Model	2	19	540	14	15
Haberman	2	3	306	13	17
Ionosphere	2	34	351	6	11
Pima	2	8	768	20	15
Vertebral 2C	2	6	310	20	14
Vertebral 3C	3	6	310	16	12

Note: # denotes "the number of object"; Pima represents "Pima Indians Diabetes". And each data has 25 experiments (five missing degrees and 5 classifiers, $5 \times 5 = 25$).

Table 5
The results of the proposed imputation in mixed noise and missing degree of data.

Dataset	#Class	#Attributes	#Samples	MAR (#win)	MCAR (#Win)
Banknote	2	5	1372	12	10
Blood	2	4	748	12	10
Climate	2	19	540	12	11
Haberman	2	3	306	7	10
Ionosphere	2	34	351	5	4
Pima	2	8	768	12	11
Vertebral 2C	2	6	310	12	6
Vertebral 3C	3	6	310	12	9

Note: # denotes the number sign, "Pima" represents the Pima Indians Diabetes dataset. There are 12 experiments in each dataset.

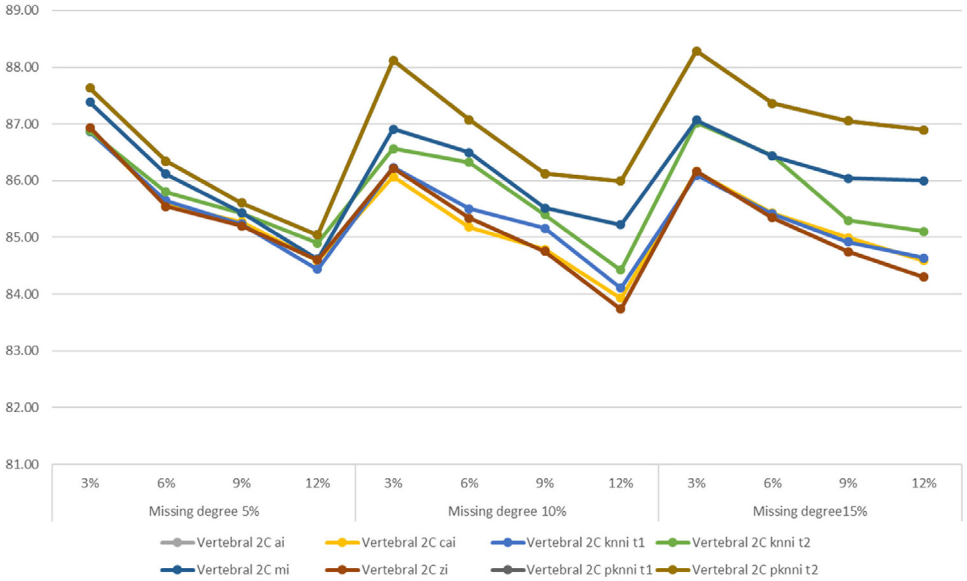


Fig. 11. The result of mixed noise and missing degrees of data for the Vertebral 2C in MAR. Note: The accuracy of all imputation methods have a larger difference when the missing degrees is ≥ 10%.

methods quickly reduced when the noise rate increased. Fig. 11 shows that the accuracy of all imputation methods have a larger difference when the missing degrees is ≥ 10%, and the proposed method is better than the listed methods for the Vertebral 2C in MAR. Fig. 12 shows that the accuracy of the different imputation methods have a larger difference in the missing degrees is 10%, and the noise rate is higher, the accuracy is reduced for the Banknote dataset in MCAR. Finally, Fig. 13 shows that the accuracy of all imputation methods have a

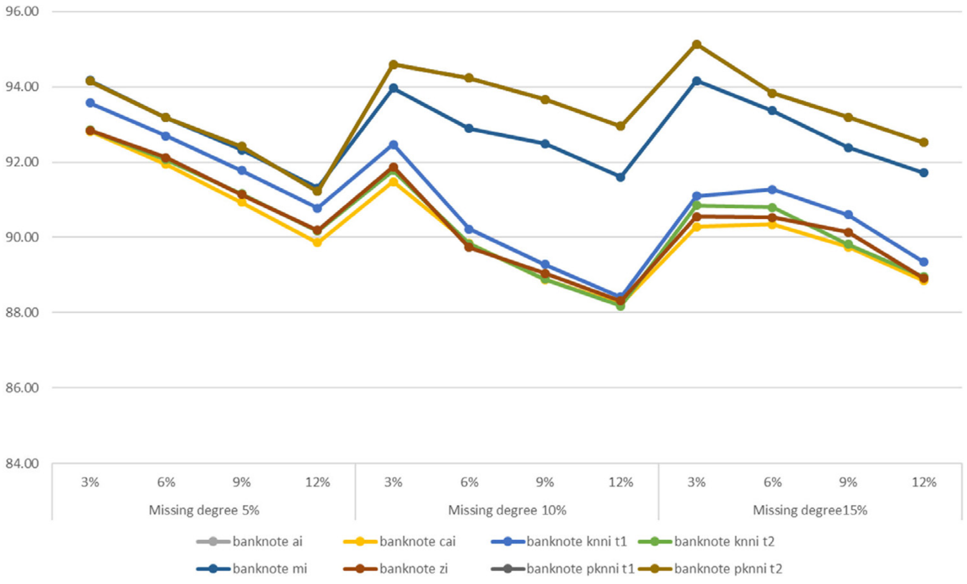


Fig. 12. The result of mixed noise and missing degrees for the Banknote dataset in MCAR. Note: the accuracy of the different imputation methods has a larger difference in the missing degrees is 10%, and the noise rate is higher, the accuracy is reduced.

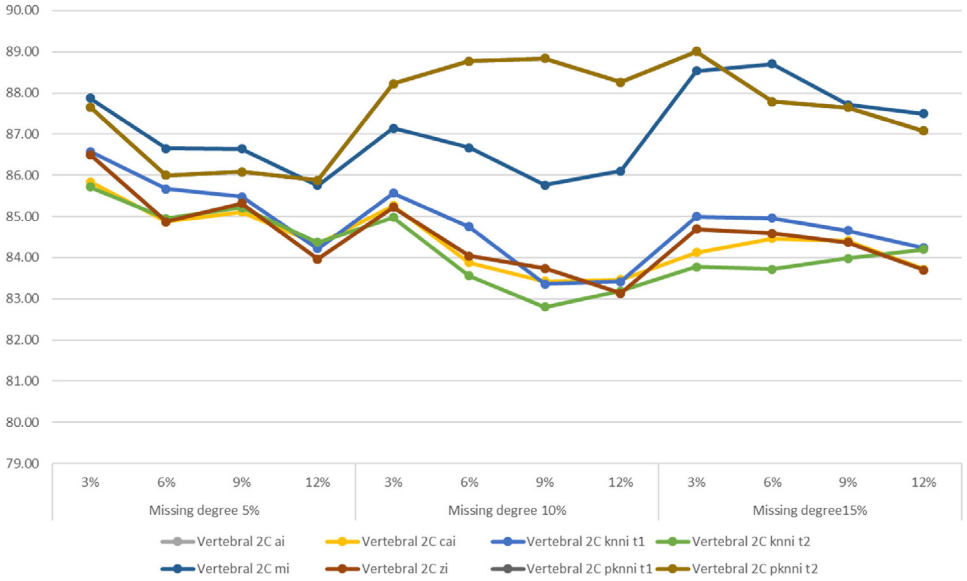


Fig. 13. The result of mixed noise and missing degree for the Vertebral 2C in MCAR. Note: the accuracy of all imputation methods have a larger fluctuation when the missing degrees is 10%.

larger fluctuation when the missing degrees is 10% for the Vertebral 2C in MCAR, however, the proposed method had good performance. In summary, the proposed imputation had better performance in different missing degrees and noise rates. It shows that the proposed imputation based on purity is a robust imputation method.

4.3. Financial distress application for MNAR experiment

In the TEJ datasets experiment, this study collected the electronic industry's quarterly financial statement data from the Taiwan Economic Journal. The data periods spanned from 1/2008 to 3/2013 quarterly with 15679 records. The practical financial statement data is important to determine the health/distress of corporations. However, the listed companies still have some concerns with revealing their information due to the economic scale, company features, and security…etc. The corporate financial statement data is similar to patient medical

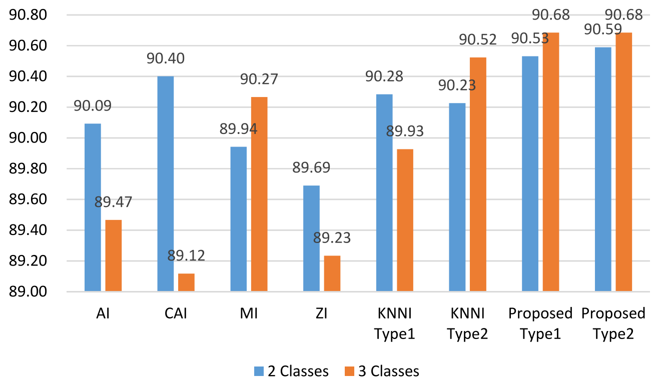


Fig. 14. The average accuracy in the two-class and three-class TEJ datasets.

Table 6
TEJ datasets description.

Attributes	Financial ratios	Formulate
X1	ROA(A)-EBI %	EBI net income/Average total assets × 100%
X2	ROA(B)-EBITDA%	(Ordinary net income + Interest expense × (1%–25%))/Average total assets × 100%
X3	ROA(C)-EBIDA%	EBIDA net income/Average total assets × 100%
X4	Return on Equity%-A	Ordinary income/Average equity × 100
X5	Operating Expense Ratio	Operating expenses/Operating net income × 100
X6	Cash Flow Ratio	Cash flow-operating/Current liabilities × 100
X7	Cash flow per share	(Cash flow-operating – Preferred stock dividends)/Weighted average equity × 10
X8	Sales per share	Operating net income/(Common stocks + Preferred stocks + Res. – Capital Increase – the number of treasury shares × 10) × 10
X9	Sales growth	(Operating net income – Last year operating net income)/ABS (Last year operating net income) × 100%
X10	Operation income growth	(Operating income – Last year operating income)/ABS(Last year operating income) × 100%
X11	Total assets growth	(Total assets – Last year total assets) BS(Last year total assets) × 100%
X12	Current ratio	Current assets/Current liabilities × 100
X13	Quick ratio	(Current assets – Inventories – Prepaid & Advance – Other current assets – Long-term investments held for disposal)/Current liabilities × 100
X14	Liabilities %	Total liabilities/Total assets × 100
X15	Equity/Total assets	Shareholders' equity/Total assets × 100
X16	Total asset turnover	Nearly four-quarter revenues/Average total assets
X17	Accounts receivables turnover	Nearly four-quarter revenues/Average Accounts receivable and Notes receivable + Notes receivable discounted)
X18	Inventory turnover	Nearly four-quarter operating costs/Average inventory
X19	Fixed asset turnover	Nearly four-quarter revenues/Average fixed assets
X20	Working capital to total assets ratio	(Current assets – Current liabilities)/Total assets
X21	EBIT to total assets ratio	EBIT/Total assets
X22	Cash flow to total liability ratio	Net cash flow-operating/Total liabilities × 100%
X23	Liquidity ratio	Current assets/Total assets
X24	Cash/Total assets	Cash/Total assets
X25	Current liabilities/Total assets	Current liabilities/Total assets
X26	Fixed Assets/Liabilities and shareholder's equity	Fixed Assets/Liabilities and shareholders' equity
X27	Shareholder's equity/Total assets	Shareholders' equity/Total assets
X28	Total liabilities	Current liabilities + Long-term liabilities + Other liabilities
X29	Per-Tax Income %	Counting operating income before tax/Operating net income × 100%
X30	Net income%	Net income before minority/Operating net income × 100%
X31	Operating gross profit margin	Gross profit/Net sales × 100%
X32	Operating profit ratio	Operating income/Net sales × 100%
X33	Return on Equity % – B	Net income/Average total shareholders' equity × 100%
X34	EPS/Total assets	EPS/Total assets
Class	Class label	Asset Earning Power = Earnings Before Taxes (EBT)/Total Assets

Table 7
The class type formulate.

Class type	Formulate
Two class	Health: Asset Earning Power > 0 Distress: Asset Earning Power ≤ 0
Three class	Health: Asset Earning Power > 0.02 Alert: $0 <$ Asset Earning Power ≤ 0.02 Distress: Asset Earning Power ≤ 0

Table 8
The accuracy of two class labels in the TEJ datasets experiment.

Method	J48	MLP	NB	BN	RF	Avg
AI	**100**	96.19	64.18	90.08	**100**	90.09
CAI	**100**	97.78	**64.19**	90.02	**100**	90.40
MI	**100**	95.36	**64.19**	90.15	**100**	89.94
ZI	**100**	94.15	**64.19**	90.09	**100**	89.69
kNNI type 1	**100**	97.10	64.18	90.13	**100**	90.28
kNNI type 2	**100**	96.80	64.17	90.15	**100**	90.22
PkNNI type 1	**100**	98.31	**64.19**	90.15	**100**	90.53
PkNNI type 2	**100**	**98.54**	64.17	**90.22**	**100**	**90.58**

Type 1 used a class label to estimated missing values, type 2 did not use a class label.

data. Therefore, the missing values of practical financial statements are considered as MNAR type. Furthermore, the original TEJ dataset contains 463 attributes, from literature and experts' suggestion, after data transform and computing financial ratio, 34 financial ratios are the main attributes (as Table 6) for building financial distress model. At last, 9299 instances with 1164 missing values remained. Table 6 shows the definition and formula of the financial ratio of the collected datasets.

In the financial distress prediction model, many researchers only use the two-class label (health or distress), therefore, this study not only examines the two-class label but also extends to a three-class label (health, alert and distress) for building a financial distress prediction model and datasets with missing values. Table 7 shows the formula for the class attribute based on the concept of this study. Next, the experiment followed the proposed procedure in Section 2. The results of the two-class and three-class are shown in Tables 8 and 9, respectively.

Table 9
The accuracy of three class labels in the TEJ datasets experiment.

Method	J48	MLP	NB	BN	RF	Avg
AI	**100**	91.92	69.34	86.05	**100**	89.46
CAI	**100**	90.01	69.40	86.17	**100**	89.11
MI	**100**	94.14	**71.23**	85.94	**100**	90.26
ZI	**100**	90.76	69.34	86.05	**100**	89.23
kNNI type 1	99.98	94.30	69.45	85.88	**100**	89.92
kNNI type 2	99.98	97.05	69.41	86.16	**100**	90.52
PkNNI type 1	**100**	**97.69**	69.53	**86.19**	**100**	**90.68**
PkNNI type 2	**100**	**97.69**	69.53	**86.19**	**100**	**90.68**

type 1 used a class label to estimated missing values, type 2 did not use a class label.

Fig. 14 shows the average of accuracy for the two-class and three-class TEJ datasets. Tables 8 and 9, show that the proposed imputation method had better accuracy in the two- and three-class datasets. The experimental results also show that handling missing values by using an imputation method is more effective than directly deleting missing values.

4.4. Finding

Based on the results of the MAR, MCAR, and MNAR experiments, the findings are as follows.

(1) Novelty of this study

(a) Less time complexity
This study proposed a novel purity-based k nearest neighbor imputation method. The proposed imputation method compared with the listing imputation methods in time complexity as follows.
SVM – big O is $O(n^3)$, where n is the number of records (Abdiansah and Wardoyo, 2015).
MLP – big O is $O(n^2)$, but it is affected by other parameters, such as the number of neurons, number of layers and structures of the neural network. Based on iterative batch-mode learning schemes to reduce the dominant cost from O(mn^2) to O(mn). (Mizutani and Dreyfus, 2001).
The proposed – big O is *O(n),* the computational complexity of the proposed imputation method had the advantage of lower computational complexity. The original big O of the proposed method is *O(nk),* where k is the parameter of knn and k is a small constant that between 3 and 9. Therefore, the big O of the proposed method was *O(n).* In the same amount of data, the proposed imputation completed the imputation task more quickly. Simultaneously, the proposed imputation is less affected by noise for the estimated values, and the results is relativity stable as Figs. 10–13.

(b) Three-class financial distress experiment
In financial distress research, this study is different from the previous studies. This paper also experiments with two-class and three-class financial distress datasets. The results show that the proposed imputation method is a good imputation method in two-class and three-class experiments as Fig. 14.

(2) MAR and MCAR experiment

The findings of the MAR and MCAR experiments have a common conclusion, which is: (1) when the noise rate rose, the accuracy declined, (2) when missing degrees increased, the accuracy decreased, and (3) in lower noise rate and missing degrees, all imputation methods were robust. Next we step by step described as follows:

(a) Proposed imputation method
Figs. 4 and 5 and Tables 4 and 5 shows the total number of successful experiments. That means the proposed imputation method can be applied to a variety of missing degrees of data and classifiers. First, in the Banknote, Blood, Pima, Vertebral 2 class and Vertebral 3 class datasets, the proposed imputation method had the best accuracy in listing classifiers and different missing degrees of data. Second, in the Blood and Haberman datasets, when the datasets had high missing degrees, the proposed imputation method achieved better results than other popular imputation techniques. Therefore, the proposed imputation method was suitable for the two datasets, which tended to contain the higher missing values. Third, as shown in Table 4, the low-dimensional datasets obtained better ranking due to the impact of purity. Finally, in the Vertebral Column datasets, it was found that the proposed method had a better imputation outcome when the same datasets were divided into more categories. This proves the advantages of the proposed method in imputing multiple class datasets.

(b) The listed imputation methods

(i) In the AI, CAI and ZI imputation method
The average accuracy decreased when the missing degrees increased. This shows the uncertainty of these imputation methods. In four popular imputation methods, CAI and ZI were relatively stable methods, because one is the class label, which produces the results of the global considerations, and the other is the normalization that outputs the similar results.

(ii) MI method
In the Climate, Haberman, Ionosphere, Pima, and Vertebral datasets, the MI imputation method is second performance (sub-optimized) but unstable. And the MI had not a good accuracy in small datasets.

(iii) The effect of the class label
In proposed imputation method, whether the class label was considered (Type 1 or Type 2), from Figs. 6–9 and Tables A and B, the proposed imputation method was not significantly affected by the class label. Furthermore, we find that there is almost no difference in the imputation performance of Type1 and Type 2. Because the proposed purity computation has re-assigned the instance into the nearest data class, this proves that the proposed method is less affected by the class label.

(c) Different missing degrees experiments
In experimental design, this study is unlike previous studies, which only explored a single missing degree. Eight different datasets and five different missing degrees were employed in the experiment to demonstrate the availability of the proposed imputation method. Figs. 4 and 5 and Table 4 show that the proposed method performed well under different missing values. It was a novel experiment with profound ramifications and it could be interesting.

(d) The experiment of mixed noise rate and missing degree
Table 5 and Tables A and B of the Appendix show that the proposed imputation method was better than the listing imputation methods in MAR and MCAR experiments, except MI in

Table A
MAR experimental results for mixed noise and missing degree.

Dataset	Noise	Missing degree 5%				Missing degree 10%				Missing degree15%			
		3%	6%	9%	12%	3%	6%	9%	12%	3%	6%	9%	12%
banknote	ai	92.17	91.38	90.97	89.78	90.28	89.28	88.59	87.68	88.52	87.96	86.93	86.24
	cai	92.23	91.23	90.79	89.62	90.14	89.10	88.28	87.45	88.32	87.72	86.58	86.10
	knni t1	92.38	91.36	90.87	89.77	90.82	89.66	88.91	87.99	88.86	88.19	86.97	86.36
	knni t2	92.38	91.47	90.91	89.72	90.79	89.48	88.67	87.66	89.26	88.64	87.19	86.57
	mi	94.05	92.94	92.63	91.28	94.41	93.04	92.38	91.26	94.28	93.43	92.10	91.72
	zi	92.21	91.37	90.98	89.78	90.28	89.27	88.59	87.68	88.52	87.96	86.93	86.24
	pknni t1	**94.37**	**93.41**	**92.77**	**91.70**	**94.94**	**93.73**	**92.95**	**91.84**	**95.00**	**94.27**	**92.96**	**92.48**
	pknni t2	**94.37**	**93.41**	**92.77**	**91.70**	**94.94**	**93.73**	**92.95**	**91.84**	**95.00**	**94.27**	**92.96**	**92.48**
blood	ai	80.54	80.67	80.60	80.95	80.14	80.29	80.46	80.59	79.98	80.21	80.58	80.40
	cai	80.34	80.43	80.30	80.94	79.88	79.85	80.68	80.70	79.98	79.84	80.50	80.44
	knni t1	80.24	80.22	80.09	80.37	79.91	79.78	80.14	80.30	79.49	79.72	79.95	80.19
	knni t2	80.40	80.52	80.43	80.76	80.82	80.87	80.69	80.86	81.93	81.71	82.04	81.51
	mi	80.81	80.94	80.67	80.84	80.95	80.69	80.70	80.72	80.91	81.49	81.67	81.52
	zi	80.54	80.67	80.60	80.95	80.14	80.29	80.46	80.59	79.98	80.21	80.58	80.40
	pknni t1	**81.31**	**81.34**	**81.29**	**81.24**	**82.34**	**82.24**	**82.50**	**82.37**	**83.41**	**83.76**	**83.95**	**83.71**
	pknni t2	**81.31**	**81.34**	**81.29**	**81.24**	**82.34**	**82.24**	**82.50**	**82.37**	**83.41**	**83.76**	**83.95**	**83.71**
climate	ai	97.60	97.29	96.54	96.39	96.82	96.58	96.00	96.16	97.27	96.90	96.46	96.37
	cai	97.53	97.18	96.46	96.39	96.83	96.58	95.99	96.11	96.89	96.78	96.34	96.01
	knni t1	97.63	97.30	96.64	96.60	97.02	96.52	96.10	96.15	97.29	97.02	96.73	96.24
	knni t2	97.56	97.29	96.74	96.57	97.15	96.85	96.48	96.42	97.66	97.42	97.10	97.17
	mi	97.53	97.19	96.50	96.61	97.57	97.06	97.15	96.92	97.86	97.30	97.24	96.67
	zi	97.60	97.29	96.54	96.39	96.82	96.58	96.00	96.16	97.27	96.90	96.46	96.37
	pknni t1	**97.76**	**97.53**	**97.05**	**97.02**	**98.26**	**98.01**	**97.91**	**97.73**	**98.61**	**98.39**	**98.31**	**98.17**
	pknni t2	**97.76**	**97.53**	**97.05**	**97.02**	**98.26**	**98.01**	**97.91**	**97.73**	**98.61**	**98.39**	**98.31**	**98.17**
Haberman	ai	79.39	79.48	78.98	79.21	79.14	78.86	78.89	79.08	79.29	78.29	78.26	78.82
	cai	79.36	79.59	78.99	79.20	79.13	78.85	78.87	79.09	79.29	78.30	78.26	78.82
	knni t1	79.46	79.59	78.98	79.12	79.15	79.12	79.05	79.24	79.19	78.45	78.45	78.98
	knni t2	79.22	79.45	78.92	79.25	79.14	78.98	79.19	79.08	78.88	78.48	78.37	79.02
	mi	79.55	79.65	**79.16**	79.35	**79.93**	**79.42**	79.48	79.42	**80.01**	79.07	**79.11**	79.61
	zi	79.37	79.47	78.99	79.21	79.14	78.86	78.88	79.08	79.29	78.29	78.26	78.82
	pknni t1	**79.67**	**79.72**	79.00	**79.40**	79.48	79.28	**79.54**	**79.48**	79.96	**79.09**	78.91	**79.75**
	pknni t2	**79.67**	**79.72**	79.00	**79.40**	79.48	79.28	**79.54**	**79.48**	79.96	**79.09**	78.91	**79.75**
Ionosphere	ai	94.18	94.44	94.13	93.99	94.28	93.87	93.91	93.84	94.30	94.42	94.26	94.43
	cai	94.21	94.43	94.18	94.01	94.27	93.77	93.93	93.93	94.26	94.32	94.20	94.38
	knni t1	94.41	94.42	94.50	94.38	94.64	94.34	94.52	94.26	94.67	94.83	94.68	94.70
	knni t2	94.44	94.59	94.44	94.33	94.47	94.32	94.36	94.28	94.86	95.01	94.81	95.11
	mi	**94.92**	**94.91**	**94.95**	**95.03**	95.54	**95.95**	95.58	**95.89**	96.01	**96.46**	96.33	96.33
	zi	94.18	94.44	94.13	93.99	94.28	93.87	93.91	93.84	94.30	94.42	94.26	94.43
	pknni t1	94.67	94.83	94.64	94.72	**95.78**	95.62	**95.68**	95.56	**96.15**	96.33	**96.43**	**96.75**
	pknni t2	94.67	94.83	94.64	94.72	**95.78**	95.62	**95.68**	95.56	**96.15**	96.33	**96.43**	**96.75**
pima	ai	82.61	81.67	80.62	81.10	81.77	81.61	80.63	80.24	81.76	80.79	79.48	80.09
	cai	82.39	81.72	80.68	81.05	81.76	81.53	80.58	80.24	81.47	80.86	79.86	80.10
	knni t1	82.89	82.04	81.14	81.39	82.03	82.03	80.84	80.54	81.85	81.26	80.62	80.26
	knni t2	82.67	82.50	81.13	81.64	82.42	81.98	81.16	80.48	82.23	81.56	80.73	80.82
	mi	83.22	82.85	81.87	82.13	83.51	83.01	82.13	81.67	84.00	82.96	82.38	82.35
	zi	82.61	81.67	80.62	81.10	81.77	81.61	80.63	80.24	81.76	80.79	79.48	80.09
	pknni t1	**83.91**	**83.61**	**82.57**	**83.11**	**85.45**	**84.67**	**84.11**	**84.02**	**87.03**	**86.45**	**85.86**	**86.01**
	pknni t2	**83.91**	**83.61**	**82.57**	**83.11**	**85.45**	**84.67**	**84.11**	**84.02**	**87.03**	**86.45**	**85.86**	**86.01**
Vertebral 2C	ai	86.94	85.55	85.20	84.61	86.22	85.34	84.75	83.74	86.17	85.35	84.75	84.30
	cai	86.90	85.57	85.28	84.58	86.07	85.19	84.79	83.93	86.16	85.43	84.99	84.59
	knni t1	86.86	85.65	85.25	84.44	86.23	85.50	85.16	84.12	86.10	85.42	84.92	84.64
	knni t2	86.88	85.82	85.43	84.90	86.57	86.32	85.40	84.43	87.02	86.45	85.30	85.11
	mi	87.38	86.12	85.44	84.62	86.91	86.49	85.52	85.22	87.07	86.43	86.05	86.02
	zi	86.94	85.55	85.20	84.61	86.22	85.34	84.75	83.74	86.17	85.35	84.75	84.32
	pknni t1	**87.63**	**86.35**	**85.62**	**85.05**	**88.12**	**87.08**	**86.12**	**86.01**	**88.29**	**87.36**	**87.07**	**86.91**
	pknni t2	**87.63**	**86.35**	**85.62**	**85.05**	**88.12**	**87.08**	**86.12**	**86.01**	**88.29**	**87.36**	**87.07**	**86.91**
Vertebral 3C	ai	86.48	85.82	85.25	83.42	85.65	85.53	84.51	83.53	85.97	85.04	84.58	83.55
	cai	86.48	85.97	85.43	83.66	85.71	85.21	84.39	83.29	85.61	84.82	83.82	83.39
	knni t1	86.21	86.22	85.77	84.21	85.55	85.37	84.42	84.25	85.78	85.27	84.68	83.64
	knni t2	87.23	86.53	85.92	84.31	86.92	86.32	85.32	84.86	88.07	86.89	86.23	85.48
	mi	86.96	86.34	85.68	84.30	85.97	85.83	85.31	85.24	86.69	85.76	85.55	85.26
	zi	86.48	85.82	85.25	83.42	85.65	85.53	84.51	83.53	85.97	85.04	84.58	83.55
	pknni t1	**88.12**	**87.40**	**87.01**	**85.41**	**87.98**	**87.64**	**86.26**	**85.91**	**89.12**	**88.31**	**87.97**	**87.57**
	pknni t2	**88.12**	**87.40**	**87.01**	**85.41**	**87.98**	**87.64**	**86.26**	**85.91**	**89.12**	**88.31**	**87.97**	**87.57**

the Ionosphere dataset. From Eq. 2.6 in Rubin (1996), despite integrating the imputed values, multiple imputation only corrects the standard errors because it assumes that there is no bias. Therefore, this paper calculated the standard errors for each attribute of the Ionosphere dataset. The 34 attributes had standard errors of 32 attributes $\sigma \approx 0.5$, showing MI to good accuracy.

We found that all accuracies of PKNNIt1 were the same as PKNNIt2 in all mixed experiments. Hence, we concluded the better performance of the proposed imputation because the proposed method is less affected by the noise.

Table B
MCAR experimental results for mixed noise and missing degree.

Dataset	Noise	Missing degree 5%				Missing degree 10%				Missing degree15%			
		3%	6%	9%	12%	3%	6%	9%	12%	3%	6%	9%	12%
banknote	ai	92.85	92.12	91.14	90.19	91.87	89.74	89.05	88.31	90.55	90.53	90.14	88.91
	cai	92.82	91.94	90.94	89.86	91.48	89.84	88.88	88.23	90.28	90.35	89.74	88.85
	knni t1	93.58	92.70	91.79	90.78	92.47	90.22	89.28	88.41	91.10	91.27	90.61	89.34
	knni t2	92.85	92.07	91.16	90.17	91.78	89.83	88.89	88.18	90.85	90.80	89.82	88.95
	mi	**94.18**	93.18	92.32	**91.31**	93.97	92.89	92.50	91.61	94.17	93.37	92.39	91.72
	zi	92.85	92.12	91.14	90.19	91.87	89.74	89.05	88.31	90.55	90.53	90.14	88.91
	pknni t1	94.15	**93.19**	**92.42**	91.22	**94.60**	**94.24**	**93.67**	**92.96**	**95.14**	**93.84**	**93.20**	**92.53**
	pknni t2	94.15	**93.19**	**92.42**	91.22	**94.60**	**94.24**	**93.67**	**92.96**	**95.14**	**93.84**	**93.20**	**92.53**
blood	ai	80.71	80.83	80.99	80.83	80.98	80.89	80.64	80.90	80.75	80.86	80.85	80.84
	cai	80.90	80.85	80.88	80.76	80.89	80.83	80.49	80.77	80.45	80.95	80.78	80.83
	knni t1	80.64	80.67	80.81	80.70	80.75	80.50	80.41	80.67	80.39	80.87	80.58	80.72
	knni t2	**81.31**	81.26	81.21	**80.98**	81.69	81.82	81.49	81.55	81.63	81.71	81.50	81.36
	mi	80.67	80.82	81.05	80.69	81.13	81.16	80.86	80.97	81.72	82.06	81.45	81.55
	zi	80.71	80.83	80.99	80.83	80.98	80.89	80.64	80.90	80.74	80.86	80.85	80.84
	pknni t1	81.04	**81.47**	**81.56**	80.96	**81.99**	**83.64**	**83.29**	**83.41**	**83.14**	**82.07**	**81.59**	**82.00**
	pknni t2	81.04	**81.47**	**81.56**	80.96	**81.99**	**83.64**	**83.29**	**83.41**	**83.14**	**82.07**	**81.59**	**82.00**
climate	ai	97.17	96.36	96.45	96.19	96.71	96.48	96.25	96.07	96.74	96.57	96.50	96.36
	cai	97.01	96.55	96.33	95.99	96.52	96.14	96.03	95.83	96.44	96.49	96.19	96.20
	knni t1	97.14	96.45	96.62	96.42	96.75	96.50	96.30	95.97	96.50	96.59	96.34	96.36
	knni t2	97.21	96.61	96.45	96.20	96.70	96.49	96.35	95.93	96.67	96.57	96.20	96.21
	mi	97.20	96.84	96.88	**96.83**	97.62	97.70	97.73	97.18	87.91	88.11	87.73	87.67
	zi	97.17	96.36	96.45	96.19	96.71	96.48	96.25	96.07	96.74	96.57	96.50	96.36
	pknni t1	**97.65**	**97.13**	**96.88**	96.47	**97.92**	**97.94**	**97.75**	**97.60**	**98.12**	**97.58**	**97.49**	**97.23**
	pknni t2	**97.65**	**97.13**	**96.88**	96.47	**97.92**	**97.94**	**97.75**	**97.60**	**98.12**	**97.58**	**97.49**	**97.23**
Haberman	ai	79.50	79.26	79.37	79.01	79.13	78.86	78.93	79.16	79.27	78.92	78.64	78.76
	cai	79.52	79.38	79.48	79.03	79.09	78.67	78.81	79.10	79.29	79.00	78.65	78.83
	knni t1	79.48	79.31	79.40	**79.21**	79.22	78.76	78.85	79.21	79.07	78.99	78.65	78.76
	knni t2	79.23	79.27	79.37	78.97	79.19	78.93	79.03	79.47	79.58	79.19	78.87	78.98
	mi	79.46	79.02	**79.73**	78.76	79.92	79.73	79.34	79.46	79.54	79.01	78.68	78.97
	zi	79.50	79.26	79.37	79.01	79.13	78.86	78.93	79.16	79.27	78.92	78.64	78.76
	pknni t1	**79.61**	**79.70**	79.71	78.95	**81.35**	**80.92**	**81.28**	**82.12**	**81.24**	**81.73**	**80.78**	**80.42**
	pknni t2	**79.61**	**79.70**	79.71	78.95	**81.35**	**80.92**	**81.28**	**82.12**	**81.24**	**81.73**	**80.78**	**80.42**
Ionosphere	ai	94.26	94.15	94.39	94.01	94.25	94.26	94.12	94.07	94.22	94.30	94.05	93.98
	cai	94.32	94.14	94.30	94.13	94.31	94.51	94.14	94.15	94.36	94.54	94.02	94.30
	knni t1	94.59	94.46	94.51	94.56	94.56	94.50	94.38	94.23	94.48	94.57	94.46	94.44
	knni t2	94.55	94.36	94.47	94.47	94.08	94.61	94.38	94.14	94.30	94.32	94.40	94.44
	mi	**95.99**	**96.11**	**96.74**	**95.65**	**99.12**	**97.05**	**98.05**	**98.25**	30.56	30.56	30.56	30.56
	zi	94.26	94.15	94.39	94.01	94.25	94.26	94.12	94.07	94.22	94.30	94.05	93.98
	pknni t1	95.03	94.92	95.14	95.04	95.80	96.40	96.31	96.44	**96.32**	**95.95**	**95.75**	**95.57**
	pknni t2	95.03	94.92	95.14	95.04	95.80	96.40	96.31	96.44	**96.32**	**95.95**	**95.75**	**95.57**
pima	ai	82.66	81.68	81.14	80.62	82.01	80.84	80.12	80.25	81.54	80.98	80.70	80.07
	cai	82.82	82.15	80.93	80.61	81.94	80.98	80.14	80.22	80.92	80.88	80.77	80.32
	knni t1	82.78	82.01	81.51	81.03	82.29	81.17	80.55	80.49	81.70	81.77	81.19	80.11
	knni t2	83.06	82.08	81.20	81.16	82.30	81.28	81.24	80.84	81.86	81.80	81.28	80.70
	mi	84.07	83.51	82.67	82.29	85.05	83.31	83.42	82.69	85.64	84.99	84.56	**84.77**
	zi	82.66	81.68	81.14	80.62	82.01	80.84	80.12	80.25	81.54	80.98	80.70	80.07
	pknni t1	**84.42**	**83.60**	**83.08**	**82.83**	**85.67**	**88.09**	**87.77**	**87.42**	**88.26**	**85.42**	**84.56**	83.67
	pknni t2	**84.42**	**83.60**	**83.08**	**82.83**	**85.67**	**88.09**	**87.77**	**87.42**	**88.26**	**85.42**	**84.56**	83.67
Vertebral 2C	ai	86.50	84.87	85.32	83.96	85.22	84.05	83.74	83.13	84.69	84.59	84.37	83.70
	cai	85.84	84.87	85.11	84.33	85.26	83.88	83.43	83.46	84.13	84.47	84.41	83.72
	knni t1	86.58	85.67	85.48	84.22	85.57	84.75	83.36	83.42	85.00	84.96	84.66	84.24
	knni t2	85.72	84.95	85.21	84.37	84.98	83.56	82.80	83.19	83.77	83.72	83.99	84.20
	mi	**87.88**	**86.66**	**86.64**	85.75	87.14	86.67	85.77	86.10	88.54	**88.70**	**87.71**	**87.49**
	zi	86.50	84.87	85.32	83.96	85.22	84.05	83.74	83.13	84.69	84.59	84.37	83.70
	pknni t1	87.65	86.00	86.08	**85.88**	**88.23**	**88.77**	**88.84**	**88.26**	**89.01**	87.79	87.65	87.08
	pknni t2	87.65	86.00	86.08	**85.88**	**88.23**	**88.77**	**88.84**	**88.26**	**89.01**	87.79	87.65	87.08
Vertebral 3C	ai	85.87	84.79	84.19	83.48	85.43	83.99	82.78	82.45	84.94	84.34	82.89	83.28
	cai	86.03	85.18	84.97	83.68	85.72	84.77	83.55	83.00	85.06	84.59	83.33	84.00
	knni t1	86.03	84.72	84.81	84.17	85.75	84.62	82.97	82.72	85.48	84.65	83.72	83.70
	knni t2	86.37	84.90	84.65	84.09	85.09	83.91	82.41	82.34	84.62	84.30	83.46	83.43
	mi	87.22	86.04	**86.24**	**86.15**	86.70	85.82	86.14	85.57	88.02	87.06	86.33	**87.04**
	zi	85.87	84.79	84.19	83.48	85.43	83.99	82.78	82.45	84.94	84.34	82.89	83.28
	pknni t1	**87.65**	**86.43**	86.08	85.34	**88.49**	**89.00**	**87.99**	**87.57**	**89.87**	**87.62**	**86.62**	86.51
	pknni t2	**87.65**	**86.43**	86.08	85.34	**88.49**	**89.00**	**87.99**	**87.57**	**89.87**	**87.62**	**86.62**	86.51

(3) MNAR experiment

From Section 4.3, the corporate financial statement data is similar to medical patient data. Therefore, it was considered as an MNAR type. The findings of the collected financial distress experiment were divided into two facets: (a) The effect of imputation, and (b) suggestions for investment.

(a) The effect of imputation

The TEJ dataset experiment proved that the proposed imputation is valid. The missing values were replaced by estimated values, correctly classified by the listing classifiers shown in Tables 8 and 9, and obtained over 90 percent accuracy.

(b) Suggestions for investor
The experimental results show that health class and alarm class can be clearly distinguished. To avoid the risk of financial distress, this study suggests that investors choose healthy companies to invest in the three-class case.

5. Conclusions

To treat the missing values problem, this study proposed a new imputation method for handling missing values, which can filter outliers and noise through a purity computation. This study implemented three types of experiments, including the use of UCI datasets in MAR and MCAR experiments and the use of TEJ datasets in MNAR experiments. The experimental results show that the proposed method can perform better in both the UCI datasets and TEJ datasets. Because the proposed imputation method considers data purity, the high purity instances are employed as an estimate candidate to improve the outcome.

In future work, many extended researches could be developed as follows:

(1) Based on feature selection, the proposed imputation method will be get better results;
(2) Applied the work of adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance (Mullick et al., 2018) to solve the problem of imbalance class with missing values;
(3) Based on feature-weighted penalty based dissimilarity (Datta et al., 2018); we can extend the related research in missing features.

Acknowledgment

We would like to thank Ministry of Science and Technology of Taiwan, this research was partially supported by the research project of Taiwan Ministry of Science and Technology (MOST 107-2221-E-224-036).

Appendix

See Tables A and B.

References

Abdiansah, A., Wardoyo, R., 2015. Time complexity analysis of support vector machines (SVM) in LibSVM. Int. J. Comput. Appl. 128 (3), 28–34.
Acuña, E., Rodriguez, C., 2004. The treatment of missing values and its effect on classifier accuracy. In: Banks, D., McMorris, F., Arabie, P., Gaul, W. (Eds.), Classification, Clustering, and Data Mining Applications. Springer, Berlin, Heidelberg, pp. 639–647.
Allison, P.D., 2002. Missing data: Quantitative applications in the social sciences. Br. J. Math. Stat. Psychol. 55 (1), 193–196.
Amiri, M., Jensen, R., 2016. Missing data imputation using fuzzy-rough methods. Neurocomputing 205, 152–164.
Andridge, R.R., Little, R.J., 2010. A review of hot deck imputation for survey non-response. Int. Stat. Rev. 78 (1), 40–64.
Batista, G.E., Monard, M.C., 2003. An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 17 (5–6), 519–533, 449-461.
Beaver, W.H., 1966. Financial Ratios as predictors of failure. J. Account. Res. 4, 71–111.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Chechik, G., Heitz, G., Elidan, G., Abbeel, P., Koller, D., 2008. Max-margin classification of data with absent features. J. Mach. Learn. Res. 9 (Jan), 1–21.
Datta, S., Bhattacharjee, S., Das, S., 2018. Clustering with missing features: a penalized dissimilarity measure based approach. Mach. Learn. 107 (12), 1987–2025.
Datta, S., Misra, D., Das, S., 2016. A feature weighted penalty based dissimilarity measure for k-nearest neighbor classification with missing features. Pattern Recognit. Lett. 80, 231–237.
Ding, Y., Song, X., Zen, Y., 2008. Forecasting financial condition of chinese listed companies based on support vector machine. Expert Syst. Appl. 34 (4), 3081–3089.
Donders, A.R.T., van der Heijden, G.J., Stijnen, T., Moons, K.G., 2006. Review: a gentle introduction to imputation of missing values. J. Clin. Epidemiol. 59 (10), 1087–1091.
García-Laencina, P.J., Sancho-Gómez, J.-L., Figueiras-Vidal, A.R., Verleysen, M., 2009. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing 72 (7), 1483–1493.
Garciarena, U., Santana, R., 2017. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Syst. Appl. 89, 52–65.
Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M., Franco, L., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. 50, 105–115.
John, G.H., Langley, P., 1995. Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp. 338–345.
Li, H., Sun, J., 2008. Ranking-order case-based reasoning for financial distress prediction. Knowl.-Based Syst. 21 (8), 868–878.
Li, H., Sun, J., Sun, B.L., 2009. Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors. Expert Syst. Appl. 36 (1), 643–659.
Lin, F., Liang, D., Chen, E., 2011. Financial ratio selection for business crisis prediction. Expert Syst. Appl. 38 (12), 15094–15102.
Liu, Z., Pan, Q., Dezert, J., Martin, A., 2016. Adaptive imputation of missing values for incomplete pattern classification. Pattern Recognit. 52, 85–95.
Mitra, S., Pal, S.K., 1995. Fuzzy multi-layer perceptron, inferencing and rule generation. IEEE Trans. Neural Netw. 6 (1), 51–63.
Mizutani, E., Dreyfus, S.E., 2001. On complexity analysis of supervised mlp-learning for algorithmic comparisons. In: Proceedings of the INNS-IEEE International Joint Conference on Neural Networks, Washington, DC, vol. 1, pp. 347–352.
Mullick, S.S., Datta, S., Das, S., 2018. Adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance. IEEE Trans. Neural Netw. 29 (11), 5713–5725.
Pearl, J., 1998. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
Quinlan, J.R., 1993. C45: Programming for Machine Learning. Morgan Kauffmann, p. 38.
Rubin, D.B., 1976. Inference and missing data. Biometrika 63 (3), 581–592.
Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc., New York.
Rubin, D.B., 1996. Multiple imputation after 18+ years. J. Amer. Statist. Assoc. 91 (434), 473–489.
Sammut, C., Webb, G. I. (Eds.), 2011. Encyclopedia of Machine Learning. Springer Science & Business Media.
Shao, J., 2000. Cold deck and ratio imputation. Surv. Methodol. 26 (1), 79–86.
Sun, J., Li, H., Huang, Q.H., He, K.Y., 2014. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. Knowl.-Based Syst. 57, 41–56.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520–525.
Tsai, C.F., Chang, F.Y., 2016. Combining instance selection for better missing value imputation. J. Syst. Softw. 122, 63–71.
Xia, J., Zhang, S., Cai, G., Li, L., Pan, Q., Yan, J., Ning, G., 2017. Adjusted weight voting algorithm for random forests in handling missing values. Pattern Recognit. 69, 52–60.
Yuan, Y., 2011. Multiple imputation using SAS software. J. Stat. Softw. 45 (6), 1–25 (online).
Zhou, L.G., Lu, D., Fujita, H., 2015. The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches. Knowl.-Based Syst. 85, 52–61.