Imputation Algorithm for Multi-view Financial Data Based on Weighted Random Forest

Jun Cao[1], Fanyu Wang[1], Zhenping Xie[1](✉), and She Song[2]

[1] College of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, Jiangsu, China
fanyu_wang@stu.jiangnan.edu.cn, xiezp@jiangnan.edu.cn

[2] Inspur Zhuoshu Big Data Industry Development Company Limited, Wuxi 214125, Jiangsu, China
songshe@inspur.com

Abstract. With the development of information technology, a large amount of multi-view data continues to emerge in the financial field. The absence of these multi-view data samples limits the research processing of financial data, while the popular single-view filling algorithm cannot handle the problem of missing multi-view data well. To address this problem, this study proposes a new filling method called Weighted Multi-view Random Forest (WMVRF), which innovatively combines feature importance to calculate view weights and enables missing filling of multi-view data by integrating the label prediction results from multiple views random forests. Several filling algorithms such as MissForest, Generative Adversarial Imputation Network, and KNN are compared on one real dataset and four multi-view public datasets (Handwritten, Webkb, 3Sources, BBCSport). The experimental results show that the proposed method reduces the normalized root mean square error (NRMSE) by 1.6% and outperforms the KNN, GAIN, and EM filling algorithms on the financial dataset compared to RF.

Keywords: missing data filling · random forest · ensemble learning · multi-view learning

1 Introduction

In the field of financial credit assessment, missing data filling is a prerequisite for all data analysis, the effect of its data filling will directly affect the enterprise credit assessment. How to deal with missing data scientifically and improve data quality is one of the research difficulties in the field of data mining nowadays. Real-life describes data in various forms and with different missing mechanisms, making the random forest-based missing-fill algorithm unable to meet the needs of all data. Research and analysis of multi-view missing data reveal that the traditional missing forest algorithm cannot effectively utilize the complementary information of multiple views to achieve missing filling of multi-view data. To solve this problem, this paper proposes a Weighted Multi-View Random Forest (WMVRF) filling method based on the random forest missing filling method combined with multi-view integrated learning.

R. Lotfi et al. (Eds.): UPRE 2023, AEBMR 253, pp. 55–70, 2023.
https://doi.org/10.2991/978-94-6463-218-7_8

2 Related Research

Since 1970, the solutions for missing information proposed at home and abroad can mainly include the following three ways.

1) Delete method: Simply remove the samples containing missing object data to obtain the complete dataset. And when the data samples in the dataset are small and there are more samples with missing attribute values, it will seriously affect the information content of the dataset.
2) Special value method: The missing attribute value is treated as a special attribute value different from any other attribute value, thus making the missing data set a complete data set.
3) Filling method: The prediction of missing data is performed by the trained model, and the predicted values are used to replace the missing data to get the data containing complete information, which is also the most commonly used method.

Currently, missing fill models are mainly classified into discriminative models and generative models. The discriminative model is mainly based on machine learning algorithms that train the model by using complete information to make predictions for missing values. Generative models are generally based on deep learning, using neural network models to simulate the dataset to generate simulated data closer to the real situation to fill in the missing data.

In a study for generative models to fill missing data, Yoon et al. [1] proposed Generative Adversarial Imputation Nets (GAIN) in combination with generative adversarial networks to fill missing components by generating data based on actual observations by generators, and then distinguishing whether the data are filled or true values by discriminators on the complete data after filling. With the generator and discriminator working against each other, the generator can generate the data distribution that is closest to the real data. Wang et al. [2] proposed a new unsupervised method to fill in missing data called Pseudo-label conditional generative adversarial imputation networks (PC-GAIN) based on GAIN, which utilizes potential category information to further enhance the interpolation capability and utilizes synthetic pseudo-label assisted classifiers to help the generators generate higher quality filling results.

Meanwhile, for the discriminative filling model, Stekhoven et al. [3] proposed a random forest missing filling algorithm (MissForest, RF), which is based on the random forest algorithm and can fill mixed data well. Van Buuren et al. [4] proposed Multivariate Imputation by Chained Equations (MICE), which is a repetitive simulation-based method for handling missing values, generating a set of data-complete datasets from a dataset containing missing values, where each complete dataset is generated by interpolating the missing data from the original data. Dixon et al. [5] proposed the K Nearest Neighbor Imputation (KNN) algorithm, which finds the most similar sample to the missing data sample in the dataset, and then uses the corresponding attribute values of this sample to fill the missing values.

With the complexity of data presentation, financial data are mostly characterized by polymorphism, multi-source, and multi-descriptive characteristics, and these characteristic data obtained from different ways or different levels for the same object are called multi-view data. Multi-view learning is a new machine learning method that uses multiple-view representations of things for modeling solutions, which generally needs to follow the principles of consistency and complementarity [6]. In recent years, multi-view learning has attracted extensive attention and research at home and abroad. Qiu et al. [7] proposed the entropy-weighted multi-view K-mean (EWKKM) algorithm for the multi-view clustering problem based on viewpoint weighting, which reduces the influence of noisy views or irrelevant views on multi-view clustering by assigning a reasonable weight to each viewpoint, and then improves the accuracy of clustering. Yang [8] proposed a viewpoint compatibility-based complementation method to obtain shared representations by reconstructing errors in the shared subspace of multiple views, based on which accurate complementation of multi-viewpoint data is achieved by multiple linear regression. Cano [9] creatively combine multi-view learning with ensemble learning by proposing a multi-view ensemble method that seeks consensus among weighted classes of predictions to exploit complementary information from multiple views, and the integration employs a voting scheme that weights the predictions of each view based on the training error of the classifier for views that have low precision classifiers or provide irrelevant noise information to reduce the impact of data in those views.

In terms of multi-view filling, missing data has been a difficult problem in multi-view data analysis, and the emergence of multi-view data has brought new ideas to fill the data. Multi-view missing data filling has become a current research hotspot in the field of machine learning to achieve a deep and comprehensive filling of missing multi-view data by reasonably utilizing the widely available multi-view information of the same object to improve the generalization, prediction accuracy, and robustness of a single view [10]. Shang et al. [11] proposed a new method called View Imputation with Generative Adversarial Networks, VIGAN) to fill the missing views by generative adversarial networks for the problem of missing views in multi-view data, which is based on Denoising Auto-Encoder (DAE), which outputs the reconstructed missing views from the GAN based on the pairwise data between views, and returns the missing views through the joint optimization of DAE and GAN, however, VIAGN as a composite neural network cannot handle more than two views. Zhang et al. [12] proposed Cross Partial Multi-View Networks (CPM-Nets) to comprehensively encode information from different views into clustered structured shared representations, while allowing flexible integration to handle arbitrary view missing cases, however, CPM-Nets are mainly intended for multi-view learning tasks and can obtain complete view shared subspaces, but cannot fill in the missing original data well.

The random forest algorithm, first proposed by Breiman [13], is an integrated algorithm based on decision trees that uses Bagging self-sampling to integrate multiple weak classifiers and generalize the overall model results by voting or averaging to give them higher accuracy and generality. Random forests have a wide range of practical applications, including corporate credit assessment, identification of financial statement fraud [14], and analysis of market behavior [15]. Random forest is based on decision trees, which train, classify and predict the sample data by integrating the prediction results of multiple decision trees, and also give the importance score of each variable and evaluate the role played by each variable in the classification. In terms of Multi-View Random Forest, Birant [16] proposed a Multi-View Rank-based Random Forest (MVRRF) algorithm for classifying eSports tournament results, which proposes to calculate the feature importance of each viewpoint to modify the random forest and reduce redundant and irrelevant features by selecting the top-ranked features to reduce the error while reducing the complexity of the model. Tian [17] et al. proposed a multi-view text classification method based on random forest, which effectively combines two text representation methods based on words and LDA topics, and effectively improves the text classification performance. The aforementioned studies mainly focus on the multi-view random forest for classification, which cannot solve the problem of missing multi-view data that exists in large quantities in the field of the financial credit assessment. Therefore, the paper proposes a multi-view random forest missing data filling method WMVRF based on view weighting.

3 Random Forest Filling Algorithm

Before filling the data with the random forest model, we need to deal with the missing dataset, including selecting the prediction labels and dividing the training and testing sets.

Suppose X is a data matrix of $n \times p$, and x_s is a missing variable in the matrix X. x_s is chosen as the prediction label. When selecting the label, it is generally chosen from lowest to highest based on the degree of missingness. The training set and test set are then divided according to the absence of the label x_s. As shown in Fig. 1, the final data can be divided into four parts.

The first part is the complete part corresponding to the label x_s represented by $y_{obs}^{(s)}$, while the second part is the missing part of the label x_s represented by $y_{miss}^{(s)}$. The third part is the rest of the sample in which $y_{obs}^{(s)}$ is located, and the last part is the rest of the sample in which $y_{miss}^{(s)}$ is located. $y_{obs}^{(s)}$ and $x_{obs}^{(s)}$ are used as training sets for model training, and $y_{miss}^{(s)}$ and $x_{miss}^{(s)}$ are used as test sets, and the trained model is used to predict $y_{miss}^{(s)}$ by continuously selecting the missing variables as labels until the original matrix X is complete and no longer missing.

The construction of a random forest can be broadly divided into four parts, which are random sample sampling, random feature selection, basic classifier construction, and voting mechanism. First, m samples are randomly resampled from the original training set using the Bootstrap self-service resampling method, and a total of *n_tree* samples are resampled to generate *n_tree* training sets. For each of the *n_tree* training sets, *n_tree*

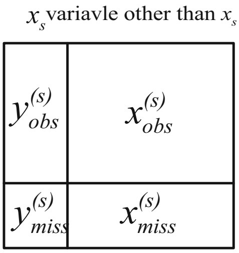

Fig. 1. Single-view dataset division.

decision tree models are trained. Second, for a single decision tree model, assuming that the number of training sample features is n, then the best feature is selected for each split based on the information gained. Then, each tree keeps splitting in this way until all training samples of that node belong to the same class. Finally, the generated multiple decision trees are formed into a random forest. For the classification problem, the final classification result is determined by the vote of the multi-tree classifier; for the regression problem, the final prediction result is determined by the mean of the multi-tree prediction values.

4 Multi-view Random Forest Filling Algorithm

4.1 Model Framework

A weighted Multi-view Random Forest algorithm (WMVRF) is proposed to address the problem that the traditional single-view filling model is only applicable to the filling of a single view while the filling accuracy is not high.

The WMVRF algorithm consists of four main components: label selection, initial filling, viewpoint weight calculation, and multi-view data integration.

The first step is to describe the data objects more comprehensively and accurately by dividing the multi-view dataset. In multi-view data, each view is sufficient for the task of missing data prediction. In multi-view data, each view is sufficient for the task of missing data prediction. Because information from different views can often complement each other, the WMVRF algorithm integrates different predictions within multiple views by calculating view weights, allowing for more comprehensive information and more accurate final prediction results.

After all the missing attributes are filled in predictively as labels, the missing data can be filled in completely. Meanwhile, the complete data after filling can be used as the initial filling for the next round of filling, and the exact filling can be achieved by continuously iterating until the filling error converges.

The general flow framework of the WMVRF algorithm is shown in Fig. 2.

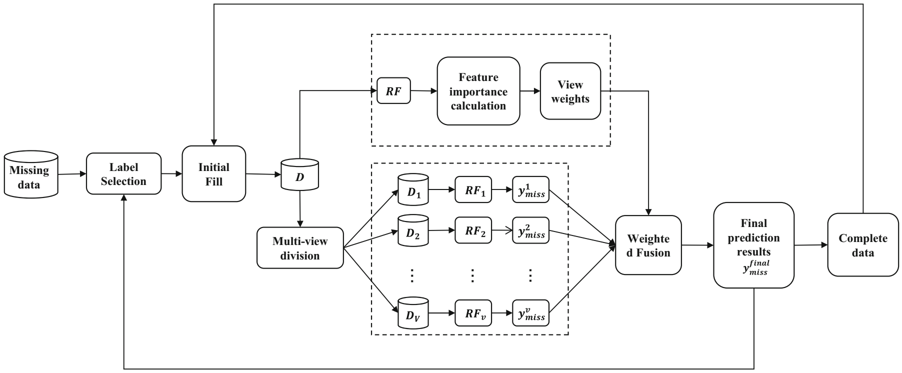

Fig. 2. An overview of the proposed WMVRF.

4.2 Algorithm Description

WMVRF is a Multi-view Ensemble learning algorithm, which reduces the variability of the prediction results of each view label by assigning different weights to each view and weighting the prediction results together to obtain a result that combines the information from multiple views.

First, before filling the multi-view missing data, the data need to be pre-processed to select the prediction labels and divide the training set and testing set, generally, the missing attributes are selected as the labels. The more complete data, the more information the decision tree classifier can learn to obtain, and thus the more accurate the prediction results. Therefore, based on the degree of missing attributes, the missing attributes are generally selected as labels in ascending order, starting from the attributes with the least degree of missingness until the data is filled in completely.

Next, the training and testing sets are divided according to the missing labels, where the samples corresponding to the complete part of the labels are used as the training set and the samples corresponding to the missing part of the labels are used as the testing set, and the trained model is used to predict the missing part of the labels. Since the dataset itself is missing, the mean filling is generally adopted as the initial filling to achieve the completion of the unlabeled part before model training.

Noteworthy, compared to single-view data segmentation, multi-view data segmentation is different in that it needs to be further broken down by the number of views. As shown in Fig. 3, suppose dataset D contains v views, and after multi-view data partitioning, the dataset D can be divided into $D_1, D_2, ..., D_v$ for x^v_{obs} and x^v_{miss} can be further subdivided into $x^1_{obs}, x^2_{obs}, ..., x^v_{obs}$, and $x^1_{miss}, x^2_{miss}, ..., x^v_{miss}$. After each view predicts the missing data, we can get the prediction value of each view on the labels $y^1_{miss}, y^2_{miss}, ..., y^V_{miss}$. After obtaining the predicted values of each view on the labels, WMVRF calculates the view weights based on the feature importance to weight the fusion of the view information to obtain the final prediction results to achieve the missing data completion.

The paper compares two weighting metrics, feature relevance, and feature importance. Feature relevance is generally based on cosine distance to calculate the correlation between label attributes and each of the view attributes, where a higher correlation is generally given a higher weight and vice versa with a lower weight. Feature importance, on the other hand, is generally based on the Gini index, which calculates the importance of the views in the overall model relative to the label predictions, in other words, it looks at the size of the overall contribution of the features within each view to each tree in the

x_s	variavle other than x_s			
$y^{(s)}_{obs}$	x^1_{obs}	x^v_{obs}	...	x^v_{obs}
$y^{(s)}_{miss}$	x^1_{miss}	x^2_{miss}	...	x^v_{miss}

Fig. 3. Multi-view dataset division.

random forest. After comparing the experimental results, it is proved that using feature importance as the weight indicator of fused perspective information to fill the missing data can get a better filling effect, so this paper chooses to use feature importance as the weight indicator for Multi-view Ensemble learning.

In this paper, we use the Gini index as an evaluation index to measure the importance of the calculated features. First, the contribution made by each feature on each decision tree is obtained by calculating the difference between the Gini index of the feature at a node, before and after branching, and then the same method is used to find the contribution values of other features. Finally, the normalized contribution of a feature is the feature importance, which is calculated by dividing the change in the Gini index of a feature by the change in the Gini index of all features, as follows.

Suppose, there are k categories and p_k denotes the weight that the k th category occupies in the node, it follows that:

$$Gini(p) = \sum_{k=1}^{k} p_k(1 - p_k) = 1 - \sum_{k=1}^{k} p_k^2 \tag{1}$$

And, for feature j, the amount of change in its Gini index at node m can be found by calculating the difference between the Gini index of the feature before branching and the Gini index after the branching at that node by Eq. (2).

$$VIM_{jm} = GI_m - GI_l - GI_r \tag{2}$$

VIM_{jm} denotes the change value of the Gini index of feature j at node m. GI_m denotes the Gini index before branching, and GI_l and GI_r are the Gini indices of the two new nodes generated after the branching of node m.

Let the set of nodes be M. From Eq. (3), we can further find the amount of change of the Gini index of feature j on the i th decision tree.

$$VIM_{ij} = \sum_{m \epsilon M} VIM_{jm} \tag{3}$$

Assuming that there are n decision trees in the random forest, the total Gini index variation of feature j is obtained from Eq. (4).

$$VIM_i = \sum_{i=1}^{n} VIM_{ij} \tag{4}$$

The feature importance $FI(j)$ of feature j can be obtained by normalizing the contribution of feature j by Eq. (5).

$$FI(j) = VIM_j (\sum_{i=1}^{c} VIM_i)^{-1} \tag{5}$$

Suppose the weight of the view is $W = \{W_V | W_1, W_2, ..., W_v\}$, for the view v, there are m features in view v. By Eq. (6), the weight W_v of view v can be calculated, which is the predicted importance of view v on the missing labels, and further the view importance of each view W_1, W_2, ..., W_V on the predicted labels x_s can be obtained.

$$W_v = \sum\nolimits_{i=1}^{m_1} FI(i) \tag{6}$$

Finally, the prediction result y_{miss}^{final} after multi-view integration can be calculated by Eqs. (7).

$$y_{miss}^{final} = \sum\nolimits_{i=1}^{V} y_{miss}^{i} W_i \tag{7}$$

The missing labels are filled sequentially until the multi-view dataset is filled. Then, the iterative optimization is started and the filled complete dataset is used as the initial filling for the next filling model, and the filling process is repeated until the stopping criterion γ is satisfied, then the difference between the filling error before and after is less than the threshold value to reach convergence, and the best filling result can be obtained.

$$\gamma_n = |\mathrm{iter}_n - iter_{n-1}| \tag{8}$$

5 Experiment

5.1 Experimental Setup

Experimental Settings

The WMVRF filling algorithm proposed in the paper and the more popular missing data filling algorithms KNN, RF, GAIN, and Mean are compared. All the following experiments are programmed using windows 11 64-bit operating system, Intel i7-10700F CPU 2. 90 GHz, 16 GB RAM, and Python 3. 8.

Evaluation Indicators

The paper used the Normalized Root Mean Square Error (NRMSE) proposed by Dauwels et al. [18] for testing the degree of difference between the filled results and the true values, as defined in Eq. (9).

$$NRMSE = \frac{1}{x_{max} - x_{min}} (\frac{1}{m} \sum_{i=1}^{m} (x_i - x_i\prime))^{1/2} \tag{9}$$

m denotes the number of samples in the data set; x_i denotes the original value, $x_i^{'}$ denotes the filled value, and x_{max} and x_{min} are the maximum and minimum values, respectively. The run results are averaged five times to reduce the correlation bias caused by missing simulated data.

Datasets

A real dataset in the field of financial credit assessment and four multi-view public

Table 1. The basic properties of the datasets.

Dataset	Samples	View	Attributes within each view
Real Dataset	436	6	5 23 17 4 6 6
3Sources	169	3	100 100 100
BBCsport	544	2	100 100
Webkb	203	3	1703 230 230
Handwritten	2000	6	240 76 216 47 64 6

datasets, including 3Sources, Webkb, BBCSport, and Handwritten, were selected for the experiment, and the composition of each dataset is shown in Table 1.

Real Dataset: A total of 13,000 enterprise information and 166 attributes are included to evaluate the comprehensive ability of enterprises from 6 directions: management ability, repayment ability, repayment willingness, profitability, enterprise qualification, and growth ability. After data pre-processing, each direction is considered as one view, and a total of 6 views and 437 enterprise information are selected to fill the experiment.

3Sources [19]: 948 articles from three well-known online news sources were selected, where each source was considered as a view. From these, 169 articles from three views containing 3560, 3631, and 3068 dimensions of the attributes were selected, and in 100 dimensional variables from each of the three views were randomly selected as a multi-view dataset to validate the method proposed in the paper.

BBCsport [19]: contains 544 sports articles from five subject areas, corresponding to five categories: athletics, cricket, soccer, rugby, and tennis. Two views were selected from these, with dimensions 3283 and 3183, and 100-dimensional variables were selected from each of the views as a multi-view dataset to validate the filling method.

Webkb [20]: contains 203 web pages of 4 categories. Each web page is described from 3 views: the content of the page, the anchor text of the hyperlink, and the text in its title containing 1703, 230, and 230 dimensions of the attributes, respectively.

Handwritten [20]: from the UCI repository, is a dataset of images of handwritten digits from 0–9. The dataset contains 2000 samples and 6 views.

Compared Methods

Simple fill: use mean fill, 0 fill, plural fill, or median fill according to the data type of the missing data. Although the operation is simple, it ignores the relationships that exist between variables.

Expectation Maximization (EM) algorithm [21]: proposed by Dempster et al. in 1977, it is an optimization algorithm based on the theory of great likelihood estimation, which uses the existence of probabilistic dependencies between variables to estimate missing data. It is suitable for multivariate missing cases but must solve the problem that the likelihood function is difficult to express and achieve only local optimum.

K-Nearest Neighbor filling (KNN): distance measurement is used to identify k samples in the dataset that are spatially similar or close to each other. Then, these samples are used to estimate the values of missing data points. The missing values of each sample are interpolated using the average of the k-neighbors found in the dataset. The Euclidean distance is generally calculated by Eq. (10), which measures the absolute distance between two points in a multidimensional space. This method is enough to use the similarity between samples to infer the missing data, the disadvantage is that the filling effect depends on the choice of K-value and similarity measure, and the computational cost is high.

$$dist(X, Y) = \left(\sum\nolimits_{i=1}^{n} (x_i - y_i)^2\right)^{1/2} \tag{10}$$

Generative Adversarial Imputation Nets (GAIN): derived from GAN networks, where the generator is used to accurately estimate the missing data and the discriminator is used to discriminate the error between the predicted and true values, thus updating the parameters of the generator and the discriminator. According to the basic principles of the GAN network, the loss of its generator and discriminator is made to get the best result simultaneously.

Pseudo-label Conditional Generation Adversarial Imputation Network (PC-GAIN): As an unsupervised missing data filling method, PC-GAIN is a further improvement on GAIN, which mainly uses the potential category information contained in the missing data to enhance the filling results. The pseudo-labels are obtained by adding pre-training to the cluster and then using the pseudo-labels to constrain the generator during model training, thus making the data generated by the generator more accurate.

Missforest (RF): a new nonparametric filling method that solves the missing data problem by training a random forest with observed values to predict missing values. Its outstanding feature is its ability to handle mixed types of data, even in the complex case of high-dimensional data, interactions, and nonlinear data structures.

5.2 Filling Errors on the Real Dataset

To verify the filling performance of the WMVRF algorithm on real corporate finance datasets, the experiments simulate the variation of the filling performance of the WMVRF algorithm under different data missing scenarios, using NRMSE as the evaluation metric, and the final results were averaged over five times.

Table 2 shows the filling results of the WMVRF algorithm and other filling algorithms at different missing rates. The experimental results demonstrate that the WMVRF has the lowest filling error at different simulated missing rates, followed by the RF, then other algorithms such as KNN and PC-GAIN, and the EM has the worst filling effect. The filling error of the WMVRF algorithm is on average 1.6% lower than that of the original RF, and more than 30% lower than that of other filling algorithms such as KNN. Among them, the MICE filling algorithm is less stable and is less effective in filling when there is less data missing. PC-GAIN has a 16% reduction in filling error compared to GAIN on this dataset. In conclusion, the WMVRF filling algorithm has the best filling results when dealing with real multi-view missing data in the field of corporate financial credit evaluation.

Further, the paper proposes four filling strategies for RF to fill missing data in multiple views, including.

(1) RF1: Combine data from multiple views into one view for filling.

Table 2. Comparison of WMVRF with state-of-the-art methods on the real financial datasets. NRMSE under various missing rates.

Model	Missing rate								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
EM	0.2174	0.2215	0.2220	0.2208	0.2207	0.2211	0.2217	0.2220	0.2223
GAIN	0.1786	0.1740	0.1723	0.1747	0.1752	0.1743	0.1735	0.1742	0.1744
MICE	0.1722	0.1562	0.1509	0.1560	0.1564	0.1842	0.1826	0.1828	0.1844
Mean	0.1540	0.1599	0.1609	0.1605	0.1607	0.1611	0.1616	0.1617	0.1618
PC-GAIN	0.1571	0.1538	0.1548	0.1573	0.1599	0.1608	0.1622	0.1638	0.1654
KNN	0.1330	0.1435	0.1442	0.1431	0.1432	0.1444	0.1463	0.1478	0.1495
RF	0.0941	0.0975	0.1009	0.1030	0.1051	0.1076	0.1103	0.1127	0.1147
WMVRF	**0.0926**	**0.0959**	**0.0993**	**0.1017**	**0.1038**	**0.1065**	**0.1093**	**0.1116**	**0.1136**

(2) RF2: separate filling of individual views.
(3) RF3: Take the average of each view's predicted results for missing labels to fill in the missing data.
(4) WMVRF: Filling in missing data based on feature importance weighted fusion of prediction information for each view.

Figure 4 shows the variation of the filling error of the four filling strategies in different cases. It is obvious that the WMVRF algorithm with the weighted fusion strategy has the best performance in filling in each missing data case, followed by RF1 and RF2, and RF3 has the worst filling effect. It can be concluded that the strategy of differentially giving different weights to each view for weighted filling is better than the undifferentiated processing strategy of RF3, and also better than the single-view filling strategies such as RF1 and RF2.

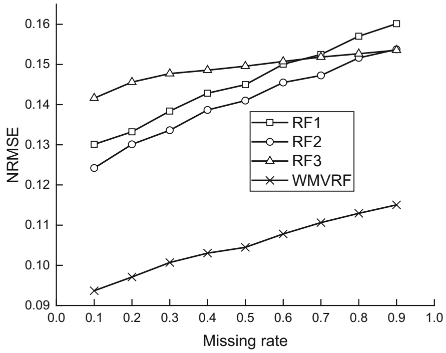

Fig. 4. Comparison of different multi-view filling strategies on real financial datasets. NRMSE with different missing rates.

Table 3. Comparison with state-of-the-art methods on a multi-view public dataset. NRMSE at 50% missing rate.

Model	Dataset			
	Handwritten	3Sources	BBCSport	Webkb
RF	0.11877	0.02314	0.10596	0.12876
WMVRF	**0.11744**	0.01779	**0.10322**	0.12772
KNN	0.16892	0.06716	0.10704	**0.10015**
GAIN	0.12788	0.36485	0.14307	0.10519
MEAN	0.29580	**0.06303**	0.10140	0.10182

5.3 Filling Errors in Public Multi-view Datasets

To further investigate the generality of WMVRF, experiments were designed on four public multi-view datasets including Handwritten, 3Source, BBCSport, and Webkb. Missing simulations were first performed on these datasets, and the missing rate was set to 50%. The experiments were compared with RF and other filling algorithms such as Mean, GAIN, and KNN as a way to demonstrate the accuracy, generality, and robustness of the WMVRF filling algorithm.

The results in Table 3 demonstrate that WMVRF fills less error than RF on all four datasets, with a 23.1% decrease in error on 3Sources, followed by 2.5% and 1.1% on BBCSport and Handwritten, and a minimum decrease of 0.8% on Webkb. It can be concluded that WMVRF improves the filling effect on the filling of multi-view data relative to RF. Comparing WMVRF with KNN, GAIN, and MEAN, it was found that WMVRF was the least effective on Handwritten and BBCSport, while MEAN and KNN were more effective on 3Sources and Webkb, respectively, demonstrating experimentally that different filling methods on different datasets have different filling effects.

5.4 Sensitivity Analysis

In the construction of a random forest by the WMVRF algorithm, the number of decision trees in the random forest is an important parameter that affects data filling. Therefore, experiments are designed on real data sets to observe the effect of different numbers of trees in the WMVRF algorithm on the filling error, and NRMSE is taken as the criterion for judging the filling error, and the final result is taken as the mean value of NRMSE after five times.

Figure 5 shows the fold change graph of the parameter simulation experiments, and the results show that the filling error shows an overall trend of gradually decreasing until convergence with the gradual increase of the number of trees in different simulated missing cases, however, the model running time increases with the number of trees running, and considering the time cost brought by increasing the number of trees, for the real data set, the WMVRF algorithm can get a better filling effect when the parameter value is 200.

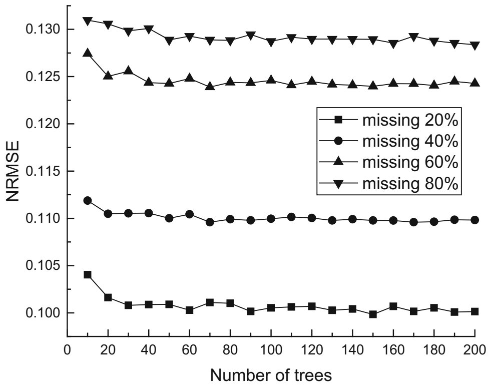

Fig. 5. NRMSE of WMVRF with different numbers of trees under various missing rates.

5.5 Iterative Optimization

The WMVRF algorithm can optimize the filling results by using the filled dataset as the initial filling for the next filling to obtain smaller filling errors, a process that can be called the iterative optimization process of the WMVRF algorithm.

Experiments are designed on real datasets with different missing cases to investigate the effect of the number of iterations of the WMVRF filling algorithm on the filling error. From Fig. 6, it can be obtained that the NRMSE decreases significantly at the first two iterations for each missing case, and then the NRMSE changes steadily as the number of iterations increases. It can be concluded that, for this real data set, the filling error can be reduced by iterating the WMVRF algorithm, and a better filling effect can be obtained at the second iteration.

To further investigate the convergence of the model iteration, a threshold value of 0.001 was set, and according to Eq. (8), γ, which is the degree of change in the root mean square error before and after, was used to measure whether the WMVRF converged after the iteration. The experimental results are shown in Fig. 7. After the 5th iteration, the value of γ is lower than the set threshold value of 0.001 in different missing conditions, so WMVRF can successfully achieve convergence after iteration.

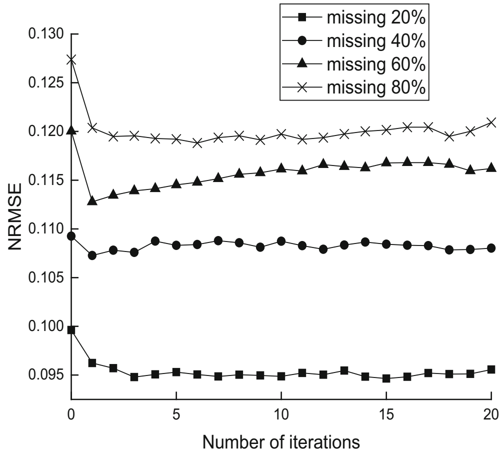

Fig. 6. The influence of the number of iterations on the NRMSE of the WMVRF.

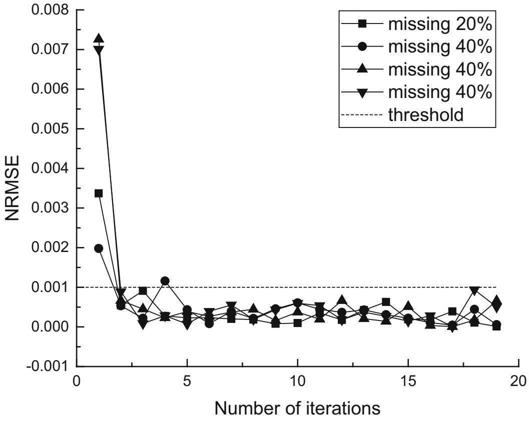

Fig. 7. Effect of the number of iterations on the difference γ of the WMVRF model.

6 Concluding Remarks

In order to solve the problem of missing and filling multi-view data in the field of financial credit evaluation, this paper proposes a filling algorithm based on multi-view ensemble learning called WMVRF, which constructs a random forest model within the view to predict missing labels and fuses the prediction information of multi-views based on the weighted importance of features, thus reducing the filling error of missing data and realizing the filling of missing multi-view data. The algorithm achieves accurate filling of missing data in multiple views.

Although the WMVRF algorithm achieves a lower filling error than RF and also has high filling accuracy on small samples and high missing data sets, it has a long processing time when dealing with high-dimensional multi-view data. Therefore, feature selection for high-dimensional datasets is needed in the future as a way to simplify the operation, while the WMVRF algorithm can be further extended to fill in multi-view missing data outside the financial credit evaluation domain.

References

1. Yoon J, Jordon J, Schaar M. Gain: Missing data imputation using generative adversarial nets[C]//International Conference on Machine Learning. PMLR, 2018: 5689–5698.
2. Wang Y, Li D, Li X, et al. PC-GAIN: Pseudo-label conditional generative adversarial imputation networks for incomplete data[J]. Neural Networks, 2021, 141: 395–403.
3. Stekhoven D J, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data[J]. Bioinformatics, 2012, 28(1): 112–118.
4. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R[J]. Journal of statistical software, 2011, 45: 1–67.
5. Dixon J K. Pattern recognition with partly missing data[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(10): 617–621.
6. Tang Jingjing, Tian Yingjie. A Review of Multi-view learning [J]. Mathematical modeling and its applications,2017,6(03):1–15+25.
7. Qiu Baozhi, He Yanfang, Shen Xiangdong. Multi-view kernel K-means algorithm based on entropy weighting [J]. Journal of Computer Applications, 2016, 36(6): 1619–1623.
8. Yang Xu, Zhu Zhenfeng, Xu Meixiang, et al. Multi-view data missing Completion [J]. Journal of Software, 2018, 29(4): 945–956.
9. Cano A. An ensemble approach to multi-view multi-instance learning[J]. Knowledge-Based Systems, 2017, 136: 46–57.
10. SUN S. A survey of multi-view machine learning[J]. Neural Computing and Applications, 2013, 23(7): 2031–2038.
11. Shang C, Palmer A, Sun J, et al. VIGAN: Missing view imputation with generative adversarial networks[C]//2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017: 766–775.
12. Zhang C, Han Z, Fu H, et al. CPM-Nets: Cross partial multi-view networks[J]. Advances in Neural Information Processing Systems, 2019, 32.
13. Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5–32.
14. An B, Suh Y. Identifying financial statement fraud with decision rules obtained from Modified Random Forest[J]. Data Technologies and Applications, 2020, 54(2): 235–255.
15. Suárez-Cetrulo A L, Cervantes A, Quintana D. Incremental market behavior classification in presence of recurring concepts[J]. Entropy, 2019, 21(1): 25.
16. Birant K U. Multi-view rank-based random forest: A new algorithm for prediction in eSports[J]. Expert Systems, 2022, 39(2): e12857.
17. Tian Baoming, Dai Xinyu, Chen Jiajun. Multi-view Text Classification Method Based on Random Forest., 2009, 23(4): 48–55.
18. Dauwels J, Garg L, Earnest A, et al. Tensor factorization for missing data imputation in medical questionnaires[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012: 2109–2112.
19. Cai H, Liu B, Xiao Y, et al. Semi-supervised multi-view clustering based on orthonormality-constrained nonnegative matrix factorization[J]. Information Sciences, 2020, 536: 171–184.
20. Wang H, Yang Y, Liu B, et al. A study of graph-based system for multi-view clustering[J]. Knowledge-Based Systems, 2019, 163: 1009–1019.
21. Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1–22.