Journal of Electronic Research and Application, 2026, Volume 10, Issue 4
http://ojs.bbwpublisher.com/index.php/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

Unconstrained Latent Factorization-Based Improved Relief-F

Wenting Chen, Ming Li*

Jiangsu Ocean University, Lianyungang, Jiangsu, China

Author to whom correspondence should be addressed.

Abstract: Feature selection is essential for dimensionality reduction on big data, but it faces considerable challenges when applied to high-dimensional and sparse datasets. To address these challenges, this paper proposes Unconstrained Latent Factorization-based Improved Relief-F (ULF-IR), a novel feature selection method tailored for such complex scenarios. The method integrates two main components: (1) a double factorization (DF)-based unconstrained latent factor model is employed to accurately reconstruct missing data without relying on pre-imputation or strict non-negativity constraints; (2) an improved Relief-F (IRelief-F) algorithm assigns reliable importance weights to features, effectively differentiating among highly similar features even in the presence of noise introduced during imputation. Comprehensive experiments on three real-world datasets show that ULF-IR consistently surpasses state-of-the-art methods in both classification accuracy and robustness, demonstrating its effectiveness as a dependable solution for feature selection on high-dimensional, incomplete data.

Keywords: Feature selection; Incomplete data; Latent factor; Relief-F

Online publication: May 21, 2026

1. Introduction

With the rapid development of information technology, large-scale data are continuously generated through complex interactions among massive entities. Although these data contain rich information, they often exhibit high dimensionality and sparsity, which may degrade model performance and hinder effective knowledge discovery in real-world applications. In particular, directly utilizing such data in downstream tasks can lead to reduced prediction accuracy, increased computational cost, and limited interpretability. Therefore, it is essential to perform feature selection to identify the most informative features, eliminate redundant and irrelevant information, and improve both efficiency and model generalization.

Existing feature selection methods can be broadly categorized into filter, wrapper, and embedded approaches. Among them, filter-based methods evaluate feature importance independently of specific learning algorithms, making them computationally efficient and scalable to high-dimensional data. In

addition, they are less prone to overfitting and can be easily applied to various tasks. However, filter-based methods struggle with sparse data, where missing values hinder accurate feature evaluation. Therefore, data imputation is necessary to support reliable feature selection.

Typical data imputation methods include statistical, temporal, machine learning, and deep learning approaches. However, statistical and temporal methods are limited in capturing complex patterns, while machine learning and deep learning methods often require sufficient data and may suffer from overfitting in sparse scenarios [1,2]. In contrast, matrix factorization effectively models latent structures and provides accurate and robust estimation of missing values, making it suitable for sparse data [3].

Consequently, extensive research has been conducted on matrix factorization methods. Given that real-world observations are typically associated with non-negative physical meanings, non-negativity constraints are often imposed on latent factor matrices to ensure the plausibility of estimated values, leading to the development of non-negative matrix factorization (NMF) [4]. However, NMF and its variants perform poorly on datasets with high missing rates, where pre-imputation of missing entries may further degrade prediction reliability. To address this issue, Luo *et al.* proposed a non-negative latent factor (NLF) model based on double factorization (DF), incorporating a single-element-dependent update scheme [5]. Nevertheless, the strategy adopted to enforce non-negativity limits the model's flexibility, scalability, and ease of implementation.

Therefore, designing a flexible latent factor model is essential for enabling reliable feature selection. In addition, considering that noise is inevitably introduced during the data imputation process, which may adversely affect feature selection, it is equally important to develop a robust feature selection method.

2. Literature review

2.1. Data imputation

Due to limitations in data collection, privacy concerns, and storage costs, real-world datasets often contain massive missing values, making accurate imputation a crucial preprocessing step. Existing estimation methods include statistical approaches such as KNN and linear regression , deep learning models including CNN and GANs, as well as matrix factorization (MF) techniques [3,6–9]. While KNN tends to reduce data variance and linear regression relies on linear assumptions, deep learning methods are computationally expensive. MF and its variants are effective but often require pre-filling of missing entries. In contrast, the double factorization-based non-negative latent factor model (DF-NLF) decomposes data into low-dimensional latent factors, offering both computational efficiency and the ability to capture latent patterns without the need for pre-filling [3,10].

Specifically, given an incomplete dataset $R \in R^{|I|\times|J|}$, the DF-NLF decompose original matrix R into the product of two non-negative d-rank latent factor matrices $A \in R^{|I|\times d}$ and $B \in R^{d\times|J|}$. The optimal factor matrices are obtained by minimizing the Euclidean distance between the observed entries and their corresponding predictions, yielding the following objective function:

$$\varepsilon(A,B)=\frac{1}{2}\sum\nolimits_{i,j\in\Lambda}\left(r_{i,j}-\sum\nolimits_{l=1}^{d}a_{i,l}b_{l,j}\right)^2, s.t.\{a_{i,l},b_{l,j}\}\geq 0; i\in\{1,\cdots,|I|\}; j\in\{1,\cdots,|J|\}. \quad (1)$$

where $a_{i,l}$ and $b_{l,j}$ denote the element in latent factor matrix A and B, respectively. Λ denotes the known elements in the original matrix R.

The objective function in **Equation (1)** is typically optimized by the Additive Gradient Descent (AGD)

algorithm, where latent factor matrices are updated according to the gradient of the loss function. However, these updates often yield negative values, violating the non-negativity constraints of the DF-NLF model. Conventional solutions either carefully tune the learning rate to limit step sizes or reset negative entries to zero after each update. Despite simplicity, both approaches have significant drawbacks: learning rate tuning is sensitive to the dataset and lacks generality, while zeroing negative values introduces bias and disrupts the gradient flow. As a result, the model becomes tightly coupled with heuristic adjustments, limiting compatibility with advanced optimization methods and potentially compromising estimation accuracy. These limitations motivate the development of a more flexible NLF formulation that naturally enforces non-negativity without relying on restrictive post-processing.

2.2. Feature selection

Feature selection is a critical preprocessing step that improves model performance and reduces computational cost by eliminating irrelevant or redundant attributes. Existing methods are typically categorized into three types: wrapper, embedded, and filter approaches. Wrapper methods identify optimal feature subsets based on predictive performance but are heavily dependent on the chosen classifier, limiting their generalizability. Embedded methods integrate feature selection within the model training process, achieving competitive results at the cost of increased computational overhead. In contrast, filter methods rank features according to intrinsic data properties independent of any specific learning algorithm, providing a favorable balance between efficiency and scalability [11]. Motivated by this, this study considers the widely-used filter-based Relief-F algorithm as a scalable and efficient approach for feature selection in large-scale datasets [12,13].

Given a complete dataset containing m samples and n features, where the feature set denoted as $F=\{f_1,f_2,\cdots,f_n\}$. For a randomly selected sample X, Relief-F identifies its k nearest neighbors: the nearest hits H_j from the same class $class(X)$, and the nearest misses $M_j(L)$ from each different class L. The weight of each feature F_i is then iteratively updated according to the following rule:

$$W(f_i)=W(f_i)-\frac{\sum_{j=1}^{k}diff\left(f_i,X,H_j\right)}{mk}+\frac{1}{mk}\sum_{L\neq class(X)}\frac{p(L)\sum_{j=1}^{k}diff\left(f_i,X,M_j(L)\right)}{1-p\left(class\left(X\right)\right)} \tag{2}$$

where $p(L)$denotes the prior probability of the class L, 1-$p(class(X))$denotes the sum of probabilities for misclassified categories, m is the number of iterations.

Although the Relief-F algorithm effectively evaluates feature importance by capturing dependencies between features and corresponding labels, it exhibits the following limitations in practice:

(1) Inability to handle sparse data, as Relief-F relies on distance-based calculations to estimate feature relevance, which become unreliable when the dataset contains many missing or unobserved entries, leading to inaccurate feature ranking;

(2) Sensitivity to noise, since the presence of irrelevant or corrupted data can distort feature weight assignments, causing redundant or non-informative features to be mistakenly prioritized.

Motivated by these issues, this work aims to develop a more robust Relief-F-based method capable of reliably identifying critical features under challenging conditions.

3. Proposed method

3.1. Unconstrained latent factorization model based on DF technique

Building on the NMF framework , the double factorization-based non-negative latent factor (DF-NLF)

model was developed to enforce non-negativity of the estimated values without requiring pre-filling of missing entries [14]. However, the non-negative constraints restricts the flexibility of this model. To overcome this limitation, this paper proposes a novel unconstrained latent factorization model based on the double factorization technique (DF-ULF). The objective function is formally defined as follows:

$$\varepsilon'(A,B)=\frac{1}{2}\sum\nolimits_{r_{i,j}\in\Lambda}\left(r_{i,j}-\sum_{l=1}^{d}\phi\left(a_{i,l}\right)\cdot\phi\left(b_{l,j}\right)\right)^2 \tag{3}$$

where ϕ denotes the single-element-dependent mapping function, which guarantee that final outputs remain non-negative.

It is important to note that the mapping function ϕ must fulfill certain conditions; otherwise, it could introduce unnecessary local optima into the original latent factor space. These required properties are formally defined as follows:

$$\forall x\in R:\begin{cases} z=\phi(x)\geq 0; \\ x=\phi^{-1}(z), ifexists; \\ \phi'(x)\neq 0. \end{cases} \tag{4}$$

Following established research, extracting latent factors from an incomplete matrix is inherently an ill-posed problem [3,14]. To mitigate this issue, the objective function in **Equation (4)** is regularized. By applying the widely used L_2-norm regularization, the objective function can be reformulated as follows:

$$\varepsilon''(A,B)=\frac{1}{2}\sum\nolimits_{i,j\in\Lambda}\left(\left(r_{i,j}-\sum_{l=1}^{d}f\left(a_{i,l}\right)f\left(b_{l,j}\right)\right)^2+\lambda\sum_{l=1}^{d}\left(f^2\left(a_{i,l}\right)+f^2\left(b_{l,j}\right)\right)\right) \tag{5}$$

where λ denotes the regularization coefficient.

With the DF-ULF model formally defined as above, the AGD algorithm is then applied to learn the latent factor matrices, enabling optimal reconstruction of the incomplete dataset.

3.2. Improved Relief-F

Building on the imputed complete data, the next step is to eliminate redundant and noisy features while preserving those that are truly informative, guided by their accurately estimated weights. Since the imputation process inevitably introduces noise, which can undermine the stability of feature selection, we incorporate a sample similarity measure into the standard Relief-F framework to mitigate the impact of noise and potential outliers. This measure captures the discrepancy between noisy samples and their typical counterparts. Specifically, we use the Pearson correlation coefficient to quantify the similarity between a given sample and its neighbors, and based on this, the weight update rule for feature $f_i\in F$ is reformulated as follows:

$$\begin{aligned} W\left(f_i\right)=W\left(f_i\right)&-\frac{1}{mk}\sum_{j=1}^{k}Sim\left(X,H_j\right)\cdot diff\left(f_i,X,H_j\right) \\ &+\frac{1}{mk}\sum\nolimits_{L\neq class(X)}\frac{p(L)\sum_{j=1}^{k}Sim\left(X,M_j(L)\right)\cdot diff\left(f_i,X,M_j(L)\right)}{1-p\left(class(X)\right)} \end{aligned} \tag{6}$$

where $Sim(X,H_j)$ and $Sim(X,M_j(L))$ represent the Pearson correlation coefficients between sample X and its nearest hitH_j, and between X and its nearest miss$M_j(L)$, respectively.

Taking $Sim(X,H_j)$ as an illustrative example, its calculation is defined as follows:

$$Pear\left(X,H_j\right)=\frac{\sum_{l=1}^{n}\left(\left(X\right)_l-\bar{X}\right)\left(\left(H_j\right)_l-\bar{H}_j\right)}{\sqrt{\sum_{l=1}^{n}\left(\left(X\right)_l-\bar{X}\right)^2}\sqrt{\sum_{l=1}^{n}\left(\left(H_j\right)_l-\bar{H}_j\right)^2}}. \tag{7}$$

where $\bar{X}$ and $\bar{H}_j$ are the average values of X and H_j, respectively.

By incorporating $Sim(X,H_j)$ and $Sim(X,M_j(L))$, the weight update process takes into account the similarity between each target sample and its neighbors. If a sample or one of its neighbors is an outlier, the corresponding Pearson correlation is low, which diminishes its impact on the feature weight adjustment. This mechanism allows the robust Relief-F algorithm to effectively mitigate the influence of noisy data points.

Based on the developments presented above, the unconstrained latent factorization-based improved Relief-F (ULF-IR) framework has been designed to handle large-scale sparse datasets. This framework integrates accurate latent factor reconstruction with robust feature weighting, enabling effective selection of informative features while mitigating the influence of noise and data sparsity.

4. Empirical analysis

4.1. General setting

To assess the effectiveness of the proposed ULF-IR, this section conducts extensive experiments on three datasets from UCI machine learning repository, as shown in **Table 1**. To simulate real-world data incompleteness, 20% of the values in each dataset are randomly removed. An 80%-20% train-test split combined with 5-fold cross-validation is adopted for model training and evaluation. For fair comparison, all datasets are scaled to a unified range of (0, 5]. Hyperparameters for each method are empirically tuned to ensure optimal performance. Each experiment is repeated 10 times, and the reported results are averaged over all runs. All experiments are conducted on a PC equipped with a 2.50 GHz processor and 8 GB RAM, with algorithms implemented in Python 3.11.

Table 1. All involved datasets

Datasets	Description	Number of features	Number of samples	Class
D1	Detect Malware Types	7107	280	2
D2	MEx Multi-modal Exercise	6262	710	7
D3	TUNADROMD	4465	241	2

4.2. Compared experiments

To evaluate the performance of ULF-IR, compared it against RSNMF and MREF [15,16]. Multiple evaluation metrics, including Accuracy and Recall, are employed to assess the feature selection methods. These metrics are obtained using two classifiers, SVM and AdaBoost, on datasets with 20% missing rate. The detailed results are summarized in **Table 2**.

Table 2. Results of compared experiments

Datasets	Method	SVM		AdaBoost	
		Accuracy	Recall	Accuracy	Recall
D1	M1	0.7771 ± 0.43%	0.8125 ± 0.62%	0.7869 ± 0.56%	0.8445 ± 0.72%
	M2	0.7925 ± 0.56%	0.8287 ± 0.13%	0.8012 ± 0.28%	0.8537 ± 0.74%
	M3	0.8092 ± 0.29%	0.8508 ± 0.53%	**0.8196 ± 0.47%**	**0.8715 ± 0.67%**
D2	M1	0.7732 ± 0.27%	0.7970 ± 0.80%	0.7890 ± 0.38%	0.8143 ± 0.54%
	M2	0.7846 ± 0.37%	0.8191 ± 0.43%	0.7935 ± 0.25%	0.8319 ± 0.68%
	M3	0.8011 ± 0.09%	0.8280 ± 0.49%	**0.8081 ± 0.35%**	**0.8438 ± 0.58%**
D3	M1	0.7828 ± 0.11%	0.8420 ± 0.21%	0.8069 ± 0.77%	0.8522 ± 0.16%
	M2	0.7974 ± 0.28%	0.8457 ± 0.49%	0.8161 ± 0.07%	0.8710 ± 0.23%
	M3	0.8104 ± 0.41%	0.8560 ± 0.85%	**0.8355 ± 0.55%**	**0.8836 ± 0.14%**

Based on the experimental results summarized in **Table 2**, M3 consistently outperforms the other methods across all datasets (D1~D3) and evaluation metrics. For instance, on the D1 dataset evaluated with AdaBoost, M3 achieves an Accuracy of 0.8196, surpassing M1 (0.7869) and M2 (0.8012) by 2.23% and 4.16%, respectively. Its Recall also reaches 0.8715, representing improvements of 2.09% over M1 (0.8537) and 3.20% over M2 (0.8445). Similar performance gains are observed on other datasets under the SVM classifier. These findings confirm that M3 (ULF-IR) is the most effective among the compared methods, delivering superior and consistent predictive performance across diverse data scenarios.

5. Conclusion

This paper proposed a novel ULF-IR algorithm to achieve accurate feature selection on high-dimensional and sparse datasets. Specifically, the approach follows a twofold paradigm: (1) a double factorization-based unconstrained latent factor model (DF-ULF) is employed to reconstruct missing entries and capture latent patterns in sparse data, and (2) an improved Relief-F algorithm (IRelief-F) is used to assign reliable feature weights while mitigating the influence of noise and redundant information. Extensive experiments on three UCI datasets demonstrate that ULF-IR consistently outperforms state-of-the-art methods in classification accuracy, validating its effectiveness as a high-performance solution for feature selection on high-dimensional, incomplete datasets. Future work will explore its application to other domains, such as image feature selection and its integration with deep learning frameworks.

Funding

National Nature Science Foundation of China (Project No.: 12401679); Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Project No.: 23KJB520006)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Fatemeh M, Ahmed H, Mohamed S, 2025, Two-Stage Hybrid Feature Selection: Integrating ACO Algorithms with a Statistical Ensemble Technique for EV Demand Prediction. IEEE Transactions on Industry Applications, 61(3): 5091–5102.

[2] Petros B, Paschalis S, Areti K, et al., 2024, Deep Learning Missing Value Imputation on Traffic Data using Self-Attention and GAN-based Methods, 2024 Panhellenic Conference on Electronics & Telecommunications (PACET), 1–4.

[3] Song Y, Li M, Luo X, et al., 2020, Improved Symmetric and Nonnegative Matrix Factorization Models for Undirected, Sparse and Large-Scaled Networks: A Triple Factorization-based Approach. IEEE Transactions on Industrial Informatics, 16(5): 3006–3017.

[4] Zhong Y, Jin L, Shang M, et al., 2020, Momentum-Incorporated Symmetric Non-Negative Latent Factor Models. IEEE Transactions on Big Data, 8(4): 1096–1106.

[5] Luo X, Zhong Y, Wang Z, et al., 2021, An Alternating-Direction-Method of Multipliers Incorporated Approach to Symmetric Non-Negative Latent Factor Analysis. IEEE Transactions on Neural Networks and Learning Systems, 34(8): 4826–4840.

[6] Sahoo A, Ghose D, 2022, Imputation of Missing Precipitation Data using KNN, SOM, RF, and FNN. Soft Computing, 26(12): 5919–5936.

[7] Yu X, Dai H, Li L, et al., 2023, Finding the Best Learning to Rank Algorithms for Effort-Aware Defect Prediction. Information and Software Technology, 2023(157): 107165.

[8] Zhao H, Li Z, He W, et al., 2024, Hierarchical Convolutional Neural Network with Knowledge Complementation for Long-Tailed Classification. ACM Transactions on Knowledge Discovery from Data, 18(6): 1–22.

[9] Luo X, Zhang X, 2025, Exploiting Defenses against GAN-based Feature Inference Attacks in Federated Learning. ACM Transactions on Knowledge Discovery from Data, 19(3): 1–20.

[10] Luo X, Zhou Y, Liu Z, et al., 2021, Fast and Accurate Non-Negative Latent Factor Analysis of High-Dimensional and Sparse Matrices in Recommender Systems. IEEE Transactions on Knowledge and Data Engineering, 35(4): 3897–3911.

[11] Xu W, Li Y, 2025, Multi-Label Feature Selection for Imbalanced Data via KNN-based Multi-Label Rough Set Theory. Information Sciences, 2025(715): 122220.

[12] Kira K, Rendell L, 1992, A Practical Approach to Feature Selection, Machine Learning Proceedings, 249–256.

[13] Kononenko I, 2005, Estimating Attributes: Analysis and Extensions of Relief-F, European Conference on Machine Learning, 171–182.

[14] Luo X, Zhong Y, Wang Z, et al., 2021, An Alternating-Direction-Method of Multipliers Incorporated Approach to Symmetric Non-Negative Latent Factor Analysis. IEEE Transactions on Neural Networks and Learning Systems, 34(8): 4826–4840.

[15] Fan H, Xue L, Song Y, et al., 2022, A Repetitive Feature Selection Method based on Improved Relief-F for Missing Data. Applied Intelligence, 52(14): 16265–16280.

[16] Takale D, Pangaonkar S, Jadhav T, et al., 2025, Enhancing Fetal Health Monitoring: Utilizing WWPA based BiGRU with Dropout Layer Regulation to Classify Fetal Health Conditions on Cardiotocography Data. Multiscale and Multidisciplinary Modeling, Experiments and Design, 2025(8): 318.