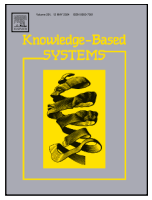
Knowledge-Based Systems 331 (2025) 114797

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Negative label-Aware and correlation-Enhanced multi-Label feature selection

Xiang Li [a,b], Huimin Fu [a,b], Xiaoou Huang [a,b], Tianyi Xie [a,b], Lingfei Ren [a,b], Wanfu Gao [c,d], Yonghao Li [a,b,*], Xin Yang [a,b]

[a] School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, 611130, China
[b] Engineering Research Center of Intelligent Finance, Ministry of Education, Southwestern University of Finance and Economics, Chengdu, 611130, China
[c] College of Computer Science and Technology, Jilin University, Changchun, 130012, China
[d] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, China

ARTICLE INFO

Keywords:
Feature selection
Multi-label learning
Label and feature correlations
Sparsity regularization
Classification

ABSTRACT

Feature selection is often seen as an essential step in multi-label learning. Existing embedded feature selection methods mainly rely on positive label space when processing the label space. However, their training objectives and learning processes almost entirely depend on the given positive label information, neglecting the valuable negative label. Moreover, the original feature space encompasses both strong and weak correlations. Unlike strong correlations, weak correlations may introduce noise that hampers the accurate representation of the feature space. To this end, we propose a feature selection method named Negative Label-Aware and Correlation-Enhanced Multi-label Feature Selection (NCMFS) that leverages information from negative label space and enhances information in feature space. Subsequently, the enriched feature space information is integrated into the reconstruction of the label space. In the NCMFS method, we further construct a sparsity regularization term based on the difference between the original weight matrix and the mirrored weight matrix to alleviate the interference of ambiguous labels and induce sparsity in the feature selection matrix. Extensive experiments on multiple datasets show that NCMFS is more efficient and provides a more competitive feature subset quality than state-of-the-art methods.

1. Introduction

Multi-label learning deals with scenarios where an individual instance is linked to multiple class labels concurrently, and the corresponding data are often described using high-dimensional feature representations[1]. Due to the abundance of invalid information in high-dimensional data, a substantial volume of data needs to be processed. Invalid information degrades model performance and thus limits the real-world applications of data mining [2,3]. Therefore, reducing the dimensionality is essential to filter out redundant or irrelevant information. Multi-label feature selection significantly reduces the dimensionality by identifying key features and eliminating redundant ones [4,5], while striving to maintain the original feature semantics [6].

In order to alleviate the dimensionality curse inherent in multi-label learning, considerable research efforts have been devoted to feature selection methods tailored for such data. In general, these approaches can be categorized into three types: filter [7], wrapper [8], and embedding methods [9]. Filter-based methods usually depend on the statistical features inherent in the data, such as mutual information, causal correlation coefficients, variance, and so on, to assess and prioritize features. As the candidate feature set evolves, the overlapping information between the features and their contribution to predicting the labels change dynamically, resulting in significant computational costs. Wrapper-based methods involve iterative generation and evaluation of feature subsets via model training, employing strategies like exhaustive search, random sampling, or heuristic search. Due to their high computational demands and the need for frequent model evaluations, these approaches are typically viable only for small-scale datasets. Embedding-based methods integrate training models using selected features [10]. This enables simultaneous optimization of both model training and feature selection. Leveraging this integrated advantage, our research specifically concentrates on embedding approaches aimed at multi-label feature selection.

Next, we delve into an analysis of several contemporary high-performing approaches. To exemplify, Gonzalez-Lopez et al. [11] design a method named GMM. GMM assesses the overall relevance of features by integrating the mutual information between them and all associated

* Corresponding author.
E-mail addresses: lixiang3270@gmail.com (X. Li), fuhuimin@swufe.edu.cn (H. Fu), xiaoouhuang507@gmail.com (X. Huang), swufejoexie@163.com (T. Xie), renlf@swufe.edu.cn (L. Ren), gaowf@jlu.edu.cn (W. Gao), liyonghao@swufe.edu.cn (Y. Li), yangxin@swufe.edu.cn (X. Yang).

https://doi.org/10.1016/j.knosys.2025.114797
Received 9 August 2025; Received in revised form 17 October 2025; Accepted 28 October 2025
Available online 2 November 2025

Table 1
The distinctions between existing approaches and our proposed NCMFS method.

Methods	Negative label information	Label correlations	Local feature correlations	Sparse and label disambiguation
GMM	×	√	×	√
FSSL	×	√	×	×
PML-ND	√	×	×	√
SSFS	×	×	√	×
LRFS	×	×	√	√
IMFS	×	√	√	√
NCMFS	√	√	√	√

Note: √indicates that the method utilizes the respective approach; ×indicates the scheme is not adopted.

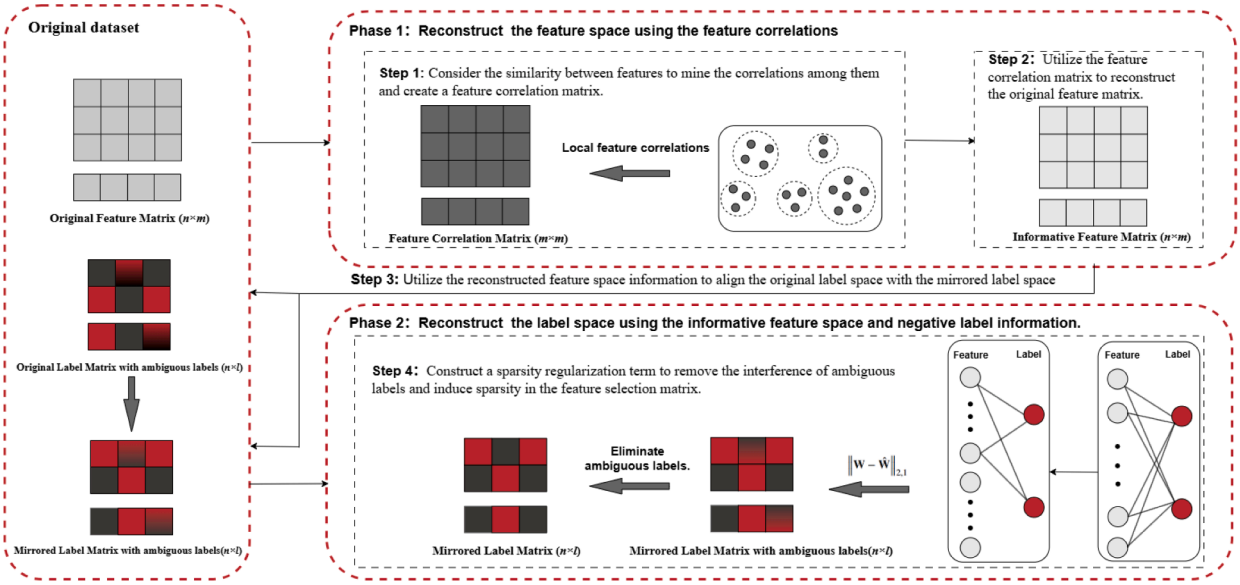


Fig. 1. The Framework of NCMFS. In phase one, a correlation matrix is established to characterize and uncover the latent clustering structure inherent in the feature set, thus enhancing the feature space by exploiting the local feature correlations. In phase two, The initial label space is converted into a mirrored counterpart through a negation operation and the reconstructed feature space information is used to align the initial label space with the mirrored label space. Furthermore, a sparsity regularization term removes the interference of ambiguous labels and induce sparsity in the feature selection matrix.

labels through geometric mean, and achieves efficient feature scoring in a distributed computing environment. However, GMM overlooks exploitable data derived from the negative label set. Liu et al. [12] design a method named FSSL. FSSL gradually optimizes the feature subset to adapt to newly added labels by dynamically selecting and fusing label-specific features. Although FSSL enriches the label space, it also leads to redundancy in the feature subset. Zhong et al. [13] propose a novel partial multi-label learning approach that effectively incorporates information from negative labels while also addressing the presence of noise in the data. However, the PML-ND method does not integrate negative and positive label information and does not exploit the correlations between features.Gao et al. [14] introduce a shared common features component designed to capture the valid information bridging the feature space and the label space. However, the SSFS method not only neglects the information contained in negative labels but also relies on a fixed graph Laplacian matrix, which may limit its adaptability. Zhang et al. [15] design a method named LRFS, By leveraging conditional mutual information, LRFS enhances the evaluation of feature-label correlations and identifies the best feature subset by integrating the analysis of feature redundancy. However, LRFS overlooks the correlations between labels and does not fully utilize the rich label information. Dai et al. [16] propose an IMFS method for feature selection in the presence of missing features and supplement implicit labels. However, the IMFS method only utilizes positive label information to supplement implicit labels and overlooks the potential of negative label information to supplement implicit labels. Table 1 presents a comprehensive comparison highlighting the distinctions between prior methods and the NCMFS approach proposed in this work.

To resolve these limitations, this paper proposes a new technique named Negative Label-Aware and Correlation-Enhanced Multi-Label Feature Selection (NCMFS) for selecting features in multi-label learning. The framework diagram of NCMFS is shown in Fig. 1. First, we construct a feature correlation matrix to characterize the inter-dependencies among features and thereby reconstruct the feature space, each feature after reconstruction contains richer information than the original. Next, by leveraging the information from the negative label space, we can more comprehensively capture the complex relationships among all labels and enhance their discriminability. Then, we utilize the reconstructed feature space information to align the initial label space with the mirrored label space based on their complementary projection matrices. Then, we construct a sparsity regularization term to remove the interference of ambiguous labels and induce sparsity in the feature selection matrix. This study offers the following significant contributions:

- We leverage the information embedded in the negative label space to capture the complex relationships among all labels and enhance their discriminability.
- We construct a feature correlation matrix to characterize the correlations among features and utilize the reconstructed feature space information to align the primitive label space with the mirrored label space.
- We construct a sparsity regularization term through analyzing the divergence between the original and mirrored weight matrix to alleviate the interference of ambiguous labels and induce sparsity in the feature selection matrix.

2. Related work

High dimensionality of features presents a significant challenge in the context of multi-label learning. Feature selection is the predominant strategy employed to tackle this problem, which extracts key features and discards redundant ones, thereby reducing dimensionality while striving to retain the semantic content of the initial features. Current techniques are generally divided into filter, wrapper, and embedded methods. Embedded approaches unify feature selection and classification into a unified learning process.

2.1. Multi-Label feature selection

A method termed MDMR was developed by Lin et al. [17], The MDMR method determines the optimal feature subset through maximizing the dependency measure that reflects the relationship between features and labels while simultaneously minimizing feature redundancy. Hu et al. [18] present a method for multi-label feature selection that effectively uncovers the common patterns existing across both the feature space and the label space. Li et al. [19] introduced the RFSFS framework, constructing an effective and adaptable sparse regularization function. Furthermore, Doquire et al. develop a feature selection method named PPT + MI [20] for multi-label learning, where the original multi-label task is transformed into a series of single-label problems to facilitate analysis. Then they use mutual information for greedy feature selection, Another method named PPT + CHI [21], PPT + CHI converts multi-label challenges into single-label ones by pruning low-frequency label combinations.

2.2. Multi-Label learning with label correlations

For multi-label feature selection purposes, effectively capturing and leveraging the dual correlations among labels and among features is paramount for achieving optimal performance. Label correlations are crucial because they reflect the intricate dependencies between different classes. Lin et al. design a method named MUCO [22], MUCO leverages the Jaccard coefficient to quantify inter-object similarity and thereby capture label correlations during streaming feature selection. Jian et al. [23] introduced MIFS, which involves learning a latent-semantic matrix with reduced dimensionality to model and represent the relationships and dependencies between different labels. Additionally, Zhang et al. [24] proposed a novel method named ROAD, which alternately alternates between adaptively discovering label correlations and performing feature selection, thereby producing a robust estimation of the dependencies among labels.

2.3. Multi-Label learning with feature correlations

Currently, feature correlations must be addressed to combat redundancy- highly covariant features provide diminishing information gains. The MSSL [25] approach proposed by Cai and Zhu reconstructs the feature manifold structure by taking into account the similarity relationships that exist between different features. Dai et al. [16] learned the synergistic interaction among features to construct feature correlations recovery matrix to recover missing features. Fan et al. [26] presented a novel scheme called LCIFS, which jointly uncovered the label correlations. They also used cosine similarity to analyze feature redundancy [27]. Although these methods leverage the correlation information between features, they fail to supplement the feature space information with such correlations, resulting in the under-utilization of the rich feature information.

In summary, exploiting label correlations and feature correlations has garnered significant attention.The previously discussed approaches improve the effectiveness of multi-label feature selection through the integration of inter-label dependencies and feature-level correlations. However, these methods only utilize information from the positive label space and fail to tap into the rich information from the negative label space. Moreover, these methods do not integrate feature correlations into label correlations. The information contained in feature correlations can provide a richer basis for the reconstructed label space, thereby achieving better feature selection results. The approach introduced in this study is specifically designed to address these limitations more effectively.

3. The proposed method

In this part, we describe the proposed NCMFS technique based on the negative label information and the feature correlation matrix to carry out feature selection work by introducing a new sparsity regularization term.

3.1. Notations

Let $X = \{x_i \mid 1 \le i \le n\} \in \mathbb{R}^{n\times d}$ represent n samples with d-dimensional feature space. $Y = \{y_i \mid \{0,1\}\} \in \mathbb{R}^{n\times l}$ be the positive label space with l labels and $\hat{Y} = \{\hat{y}_i \mid \{0,1\}\} \in \mathbb{R}^{n\times l}$ be the negative label space with l labels. $Y_{ij} = 1$ represents that X_{i*} does not belong to Y_{*j}, conversely, $Y_{ij} = 0$ represents that X_{i*} does not belong to Y_{*j}.

3.2. Feature selection based on sparse model

We posit that there exists an inherent linear dependency structure within the feature matrix X and the label matrix Y. For the purpose of revealing the connections between the feature space $X \in \mathbb{R}^{n\times d}$ and the label space $Y \in \mathbb{R}^{n\times l}$, we employ least-squares regression via the projection matrix $W \in \mathbb{R}^{d\times k}$. Consequently, a linear relational mapping is established between X and Y.

$$\min_{W\ge 0} \|XW - Y\|_F^2 \quad s.t. \quad \{W\} \ge 0, \tag{1}$$

where the components of the matrix W for feature selection can be interpreted as weights that quantify the relationship between the feature space X and the label space Y. Consequently, each element W_{ij} of W reflects the discriminative power between the i-th feature and the j-th label.

3.3. Reconstructed feature space

In reality, there is correlation between features and the value of one feature can usually be estimated from other related features. Based on this idea, we construct a feature correlation $\mathbf{C}$. Initially assigned random values, this $\mathbf{C}$ is continuously updated during the iteration process to capture richer and more accurate feature correlations. The feature correlation matrix characterize the inter-dependencies among features and thereby reconstruct the feature space. For the element X_{ij}, its reconstructed value is given by $\hat{X}_{ij} = \sum_{k=1}^{m} X_{ik}C_{kj}$, where C_{kj} signifies the dependency between the k-th and j-th features. Following this reconstruction, each feature is enriched with more information compared to its original form, which in turn enhances the learning process of the projection matrix. The component responsible for feature reconstruction is denoted as

$$\min_{\mathbf{W},\mathbf{C}} \quad \|\mathbf{XCW} - \mathbf{Y}\|_F^2 + \alpha\|\mathbf{XC} - \mathbf{X}\|_F^2 \quad s.t. \quad \{W, C\} \ge 0, \tag{2}$$

3.4. Leveraging negative label space information

To mine informative patterns concealed in the negative label space, we apply a label-wise negation over the original label matrix Y. This transformed domain is known as the mirrored label space $\hat{Y}$. Treating $\hat{Y}$ as a complementary supervision signal makes absence-related information explicit and thus easier to model. Many multi-label datasets contain far fewer positive than negative entries; $\hat{Y}$ exposes abundant positives

for the previously negative class, improving statistical efficiency and calibration. We also use least-squares regression to explore related relationships between the feature matrix $X \in \mathbb{R}^{n\times d}$ and the mirrored label matrix $\hat{Y}$ by the projection matrix $\hat{W} \in \mathbb{R}^{d\times k}$. The subsequent challenge lies in effectively combining the mirrored label space, derived from the logical negation of the original label space, with its source counterpart. Crucially, both the original and the mirrored label spaces maintain intrinsic connections to the feature space. Therefore, we can integrate them into the following form:

$$\min_{W,\,\hat{W}\geq 0} \|X\hat{W}-\hat{Y}\|_F^2+\alpha\left\|\mathbf{1}_{n\times l}-XW-X\hat{W}\right\|_F^2 \quad s.t.\ \ \{W,\hat{W}\}\geq 0, \tag{3}$$

where $\mathbf{1}_{n\times l}$ is a matrix with the same dimensions as the label matrix and all elements are 1. $X\hat{W}$ and XW denote the reconstructed mirrored label matrix and the reconstructed original label matrix, respectively. The sum of $X\hat{W}$ and XW is $\mathbf{1}_{n\times l}$ according to the negation operation. $\mathbf{1}_{n\times l}-X\hat{W}$ denotes the denoised Y.

By combining the objective functions in Formulas (2) and (3), the objective function based on the feature correlation matrix and negative label information can be formulated as:

$$\begin{aligned}\min_{\mathbf{W},\hat{\mathbf{W}},\mathbf{C}} &\left\|\mathbf{XC}\hat{\mathbf{W}}-\hat{\mathbf{Y}}\right\|_F^2+\alpha\|\mathbf{XC}-\mathbf{X}\|_F^2+\beta\left\|\mathbf{I}_{n\times l}-\mathbf{XCW}-\mathbf{XC}\hat{\mathbf{W}}\right\|_F^2\\ &s.t.\ \ \{W,\hat{W},C\}\geq 0,\end{aligned} \tag{4}$$

The original label space Y and the mirrored label space $\hat{Y}$ can be approximated by the terms XW and $X\hat{W}$. The weight values of these approximate labels are between 0 and 1. Since the feature space matrix X is fixed, the smaller the difference between W and $\hat{W}$, the less the distinction between the original labels and mirrored labels, we refer to these labels as ambiguous labels. The sparsity regularization term can ensure the row sparsity of the$\|W-\hat{W}\|_{2,1}$ term [23], which can alleviate or eliminate the impact of those similar labels in the original label space and the mirrored label space. Therefore, the sparsity regularization term based on the difference between the original weight matrix W and the mirrored weight matrix $\hat{W}$ alleviates or eliminates the impact of ambiguous labels. Thus, we arrive at the following formulation:

$$\begin{aligned}\min_{\mathbf{W},\hat{\mathbf{W}},\mathbf{C}} &\left\|\mathbf{XC}\hat{\mathbf{W}}-\hat{\mathbf{Y}}\right\|_F^2+\alpha\|\mathbf{XC}-\mathbf{X}\|_F^2+\beta\left\|\mathbf{I}_{n\times l}-\mathbf{XCW}-\mathbf{XC}\hat{\mathbf{W}}\right\|_F^2\\ &+\gamma\,\mathrm{Tr}(\mathbf{W}^\top\mathbf{LW})+\lambda\left\|\mathbf{W}-\hat{\mathbf{W}}\right\|_{2,1}\quad s.t.\ \ \{W,\hat{W},C\}\geq 0,\end{aligned} \tag{5}$$

where $L=A-S$ embodies a graph Laplacian matrix and $A\in\mathbb{R}^{d\times d}$ is an All-one matrix, The matrix S is obtained by computing the pairwise cosine distances between the rows of X. $S_{ij}=1-\frac{u_i^\top u_j}{\|u_i\|\|u_j\|}$ where u_i represents a row vector of X^T. We apply non-negative constraints to all parameter variables $\{W,\hat{W},C\}$ based on the non-negative characteristics of the labels and the subsequent update strategies.

4. Optimization

4.1. Optimization scheme

Obviously, Formula (5) is joint nonconvex w.r.t. all variables. Consequently, it is impossible to reach the globally optimal solution for this function. Furthermore, the imposition of the $l_{2,1}$-norm constraint on the feature selection term introduces a non-smooth effect. To tackle this joint non-convex issue, we propose two iterative rules. $||W-\hat{W}||_{2,1}$ is relaxed to $\mathrm{Tr}((W-\hat{W})^T D(W-\hat{W}))$, where $D\in\mathbb{R}^{d\times d}$ denotes a main diagonal matrix with its i-th diagonal unit $D_{ii}=\frac{1}{||(W-\hat{W})_{i\cdot}||+\epsilon}$, and ϵ is a vanishingly small positive constant. Therefore, we rewrite Formula (5) in the subsequent form:

$$\begin{aligned}\min_{\mathbf{W},\hat{\mathbf{W}},\mathbf{C}} &\left\|\mathbf{XC}\hat{\mathbf{W}}-\hat{\mathbf{Y}}\right\|_F^2+\alpha\|\mathbf{XC}-\mathbf{X}\|_F^2+\beta\left\|\mathbf{I}_{n\times l}-\mathbf{XCW}-\mathbf{XC}\hat{\mathbf{W}}\right\|_F^2\\ &+\gamma\,\mathrm{Tr}(\mathbf{W}^\top\mathbf{LW})+\lambda\,\mathrm{Tr}\left((W-\hat{W})^T D(W-\hat{W})\right)\quad s.t.\ \ \{W,\hat{W},C\}\geq 0,\end{aligned} \tag{6}$$

Next, we integrate the non-negative constraints of all variables $(W,\hat{W},C)$ into the objective function, and then the obtained Lagrange function is as shown below:

$$\begin{aligned}\mathcal{L}(W,\hat{W},C)=&\left\|\mathbf{XC}\hat{\mathbf{W}}-\hat{\mathbf{Y}}\right\|_F^2+\alpha\|\mathbf{XC}-\mathbf{X}\|_F^2+\beta\left\|\mathbf{I}_{n\times l}-\mathbf{XCW}-\mathbf{XC}\hat{\mathbf{W}}\right\|_F^2\\ &+\gamma\,\mathrm{Tr}(\mathbf{W}^\top\mathbf{LW})+\lambda\,\mathrm{Tr}\left((W-\hat{W})^T D(W-\hat{W})\right)-\mathrm{Tr}(\Psi W^T)\\ &-\mathrm{Tr}(\Phi\hat{W}^T)-\mathrm{Tr}(\Omega C^T)\end{aligned} \tag{7}$$

where Ψ , Φ and Ω denote the Lagrange multipliers of the corresponding W, $\hat{W}$ and C. Then, the related partial derivatives W , $\hat{W}$ and C have the following form:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial W}=&2\beta C^TX^TXCW+2\beta C^TX^TXC\hat{W}-2\beta C^TX^T\mathbf{1}+2\gamma AW\\ &-2\gamma SW\ +2\lambda DW-2\lambda D\hat{W}-\Psi.\end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\hat{W}}=&2C^TX^TXC\hat{W}-2C^TX^T\hat{Y}+2\beta C^TX^TXCW+2\beta C^TX^TXC\hat{W}\\ &-2\beta C^TX^T\mathbf{1}+2\lambda D\hat{W}-2\lambda DW-\Phi.\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial C}=&2X^TXC\hat{W}\hat{W}^T-2X^T\hat{Y}\hat{W}^T+2\alpha X^TXC-2\alpha X^TX\\ &+2\beta X^TXCW(W+\hat{W})^T+2\beta X^TXC\hat{W}(W+\hat{W})^T\\ &-2\beta X^T\mathbf{1}(W+\hat{W})^T-\Omega.\end{aligned} \tag{10}$$

Nevertheless, closed-form solutions are not available for some of the aforementioned partial derivatives. We know $\Psi_{ij}W_{ij}=0$, $\Phi_{ij}\hat{W}_{ij}=0$, and $\Omega_{ij}C_{ij}=0$, to address this challenge, we employ the Karush-Kuhn-Tucker (KKT) condition. Concretely, the formulation is obtained:

$$\begin{aligned}&[\beta C^TX^TXCW+\beta C^TX^TXC\hat{W}-\beta C^TX^T\mathbf{1}+\gamma AW\ -\gamma SW\\ &+\lambda DW-\lambda D\hat{W}]_{ij}\Psi_{ij}=0\end{aligned} \tag{11}$$

$$\begin{aligned}&[C^TX^TXC\hat{W}-C^TX^T\hat{Y}+\beta C^TX^TXCW+\beta C^TX^TXC\hat{W}-\beta C^TX^T\mathbf{1}\\ &+\lambda D\hat{W}-\lambda DW]_{ij}\Phi_{ij}=0\end{aligned} \tag{12}$$

$$\begin{aligned}&[X^TXC\hat{W}\hat{W}^T-X^T\hat{Y}\hat{W}^T+\alpha X^TXC-\alpha X^TX+\beta X^TXCW(W+\hat{W})^T\\ &+\beta X^TXC\hat{W}(W+\hat{W})^T-\beta X^T\mathbf{1}(W+\hat{W})^T]_{ij}\Omega_{ij}=0\end{aligned} \tag{13}$$

Given the objective with non-negativity constraints $W\geq 0$, $\hat{W}\geq 0$ and $C\geq 0$, we form the Lagrangian and take first-order conditions, then impose the KKT complementary-slackness conditions for non-negativity, which yield the block gradients. For each block, we split the gradient into nonnegative parts and collect terms that push the objective up into the denominator and versus down into the numerator. Therefore, W, $\hat{W}$, and C can be updated as follows:

$$W_{ij}^{t+1}\leftarrow W_{ij}^{t}\frac{(\beta C^TX^T\mathbf{1}+\gamma SW+\lambda D\hat{W})_{ij}}{(\beta C^TX^TXCW+\beta C^TX^TXC\hat{W}+\gamma AW+\lambda DW)_{ij}}, \tag{14}$$

$$\hat{W}_{ij}^{t+1}\leftarrow\hat{W}_{ij}^{t}\frac{(C^TX^T\hat{Y}+\beta C^TX^T\mathbf{1}+\lambda DW)_{ij}}{(C^TX^TXC\hat{W}+\beta C^TX^TXCW+\beta C^TX^TXC\hat{W}+\lambda D\hat{W})_{ij}}, \tag{15}$$

$$C_{ij}^{t+1}\leftarrow C_{ij}^{t}\frac{(X^T\hat{Y}\hat{W}^T+\alpha X^TX+\beta X^T\mathbf{1}(W+\hat{W})^T)_{ij}}{\begin{array}{l}(X^TXC\hat{W}\hat{W}^T+\alpha X^TXC+\beta X^TXCW(W+\hat{W})^T+\\ \beta X^TXC\hat{W}(W+\hat{W})^T)_{ij}\end{array}}, \tag{16}$$

where t indicates the current number of iterations. Algorithm 1 declares convergence when the relative change between successive objective values falls below 10^{-3} after at least three iterations, and it also imposes a maximum of 1000 iterations. To avoid trivial solutions, we add ϵ to the denominator of each update rule. The key update steps are shown in Algorithm 1.

4.2. Convergence analysis

This subsection focuses on establishing the convergence of the objective function with respect to Lee et al. [28] and Ding et al. [29].

We adopt the standard Majorization-Minimization method: first, find a function $\mathcal{G}$ that is equal to the original objective function at the current

Algorithm 1 NCMFS.

Input: Feature matrix X and label matrix Y
Parameter: Parameters α, β, γ and λ
Output: ranked feature index

1: Let $t = 0$.
2: **while** not converged **do**
3: Update the mirrored label projection matrix W according to Formula 14;
4: Update the original label projection matrix $\hat{W}$ on the basis of Formula 15;
5: Update the original label projection matrix C on the basis of Formula 16;
6: Update the relaxed diagonal matrix D on the basis of $D_{ii} = \frac{1}{||(W-\hat{W})_{i\cdot}||+\epsilon}$;
7: Update $t = t + 1$;
8: **end while**
9: **return** features on the basis of $||(W - \hat{W})_{i\cdot}||_2$.

point and is an upper bound of the original objective function everywhere within the neighborhood; then minimizing this upper bound can ensure that the original objective function is not increasing.

Definition 1. Let $\mathcal{F}(w)$ be the objective function to be minimized. Given the current point w', if the function $\mathcal{G}(w, w')$ meets the specified conditions: 1. Majorization: $\mathcal{G}(w, w') \geq \mathcal{F}(w)$ and 2. Tangency: $\mathcal{G}(w, w) = \mathcal{F}(w)$, $\mathcal{G}(w, w')$ can be understood as a supplementary function of $\mathcal{F}(w)$, where $\mathcal{G}$ and $\mathcal{F}$ signify the supplementary function and the objective function. Intuitively, $\mathcal{G}(w, w')$ is a more tractable function that majorizes the original objective function at w'; as long as the minimum of this majorizing function is taken as the new iteration point each time, the value of the original objective function will not increase. In this paper, we construct this function $\mathcal{G}$ for each variable block independently and minimize them alternately.

For any w and w', the following lemma holds:

Lemma 1. *If $\mathcal{G}(w, w')$ serves as an auxiliary function of $\mathcal{F}(w)$, then $\mathcal{F}(w)$ decreases monotonically provided the following condition holds:*

$$w^{t+1} = \arg\min_{w} \mathcal{G}(w, w^t) \tag{17}$$

Proof. According to $\mathcal{F}(w^{t+1}) \leq \mathcal{G}(w^{t+1}, w^t) \leq \mathcal{G}(w^t, w^t) = \mathcal{F}(w^t)$, Lemma 1 is proved. □

Subsequently, the crucial step is to identify a function $\mathcal{G}(w, w')$ that fulfills the aforementioned constraints. Consequently, we isolate the component of the objective that depends solely on W_{ij} and denote it by $\mathcal{F}_{ij}(W_{ij})$. By calculating the first- and second-order gradients of $\mathcal{F}_{ij}(W_{ij})$ in relation to W_{ij} yields, we have:

$$\begin{cases} F'_{ij} = (2\beta C^T X^T XCW + 2\beta C^T X^T XC\hat{W} - 2\beta C^T X^T \mathbf{1} + 2\gamma \\ \qquad AW - 2\gamma SW + 2\lambda DW - 2\lambda D\hat{W})_{ij} \\ F''_{ij} = 2\beta(C^T X^T XC)_{ii} + 2\gamma(A - S)_{ii} + 2\lambda D_{ii}, \end{cases} \tag{18}$$

Therefore, $\mathcal{F}_{ij}(W_{ij})$ is obtained:

$$\mathcal{F}_{ij}(W_{ij}) = \mathcal{F}_{ij}(W^t_{ij}) + \mathcal{F}'_{ij}(W^t_{ij})(W_{ij} - W^t_{ij}) + \frac{1}{2}\mathcal{F}''_{ij}(W^t_{ij})(W_{ij} - W^t_{ij})^2 \tag{19}$$

Now, the following function is constructed:

$$\mathcal{G}(W_{ij}, W^t_{ij}) = \mathcal{F}_{ij}(W^t_{ij}) + \mathcal{F}'_{ij}(W^t_{ij})(W_{ij} - W^t_{ij}) + \frac{(\beta C^T X^T XCW + \beta C^T X^T XC\hat{W} + \gamma AW + \lambda DW)_{ij}}{W^t_{ij}}(W_{ij} - W^t_{ij})^2 \tag{20}$$

Lemma 2. *Formula (20) serves as an auxiliary function for $\mathcal{F}_{ij}(W_{ij})$.*

Proof. First, $\mathcal{G}(W_{ij}, W^t_{ij}) = \mathcal{F}_{ij}(W_{ij})$, when $W_{ij} = W^t_{ij}$. Then, we need to prove $\mathcal{G}(W_{ij}, W^t_{ij}) \geq \mathcal{F}_{ij}(W_{ij})$ according to Definition 1, the stated inequality is equivalent to the following inequality:

$$\frac{(\beta C^T X^T XCW + \beta C^T X^T XC\hat{W} + \gamma AW + \lambda DW)_{ij}}{W^t_{ij}} \geq \beta(C^T X^T XC)_{ii} + \gamma(A)_{ii} + \lambda D_{ii} \tag{21}$$

□

The following inequalities clearly hold:

$$(\beta C^T X^T XCW)_{ij} = \beta \sum_{r=1}^{d} (C^T X^T XC)_{ir} W_{rj} \geq \beta(C^T X^T XC)_{ii} W_{ij} \tag{22}$$

$$\gamma(AW)_{ij} = \gamma \sum_{r=1}^{d} (AW)_{ir} W_{rj} \geq \gamma(A)_{ii} W_{ij} \tag{23}$$

$$(\lambda DW)_{ij} = \lambda \sum_{r=1}^{d} (DW)_{ir} W_{rj} \geq \lambda D_{ii} W_{ij} \tag{24}$$

Thus, Formula (20) is an auxiliary function of $\mathcal{F}_{ij}(W_{ij})$. Substituting this auxiliary function into Formula 17 , we obtain:

$$W^{t+1}_{ij} \leftarrow W^t_{ij} - W^t_{ij} \frac{\mathcal{F}'_{ij}(W^t_{ij})}{2(\beta C^T X^T XCW + \beta C^T X^T XC\hat{W} + \gamma AW + \lambda DW)_{ij}} \tag{25}$$

According to $\mathcal{F}'_{ij}(W^t_{ij})$ in Formula 18, the update rule takes the following form:

$$W^{t+1}_{ij} \leftarrow W^t_{ij} \frac{(\beta C^T X^T \mathbf{1} + \gamma SW + \lambda D\hat{W})_{ij}}{(\beta C^T X^T XCW + \beta C^T X^T XC\hat{W} + \gamma AW + \lambda DW)_{ij}} \tag{26}$$

Consequently, we demonstrate that the objective function exhibits non-ascending convergence with regard to W, The convergence of $\hat{W}$ and C can be proved in the same way.

5. Experiment

This section focuses on analyzing the performance of NCMFS through experiments. The experiments use ten public datasets from various domains. Among them, there is one music dataset (i.e., Emotions); one audio dataset (i.e., Birds); two image dataset (i.e., Flags and Scene); and six text classification datasets (i.e., Business, Education, Enron, Health, Recreation and Science). The specific information of the above datasets are shown in Table 2.

5.1. Experimental settings

We carry out experiments using eight advanced multi-label learning methods (MIFS [23], FLIS [30], LRFS [15], GMM [11], SSFS [14], PPT + MI [20], FSSL [12], PPT + CHI [21]) to validate the effectiveness of NCMFS in multi-label feature selection. Among them, MIFS effectively utilizes the implicit correlations between labels through low-dimensional label space decomposition and cross-label shared feature

Table 2
Details of ten benchmark datasets.

Datasets	Domain	#Instances	#Features	#Labels
Birds	Audio	645	260	20
Business	Text	5000	438	30
Education	Text	5000	550	33
Emotions	Music	593	72	6
Enron	Text	1702	1001	53
Flags	Images	194	19	7
Health	Text	5000	612	32
Recreation	Text	5000	606	22
Scene	Images	2407	294	6
Science	Text(Web)	5000	743	40

selection. The FLIS method integrates label information supplementation and the mechanisms of selected features with labels. LRFS enhances the evaluation of feature-label correlations and identifies the best feature subset by integrating the analysis of feature redundancy. GMM sesses the overall relevance of features by integrating the mutual information between them and all associated labels through geometric mean, and achieves efficient feature scoring in a distributed computing environment. SSFS designs potential structure-sharing terms and uses graph regularization technology to maintain the consistency of the structure between features and labels. PPT + MI converts multi-label problems into single-label problems and uses mutual information for greedy feature selection, which significantly reduces the problem dimension and improves the performance of classifiers. FSSL gradually optimizes the feature subset to adapt to newly added labels by dynamically selecting and fusing label-specific features. PPT + CHI converts multi-label challenges into single-label challenges by pruning low-frequency label combinations, thereby simplifying model training and improving classification results. Due to the absence of existing methods, this study uses the weight matrices of comparative algorithms to characterize differences in feature importance, and a ten-fold cross-validation scheme is uniformly adopted in NCMFS and comparative methods. All experiments were carried out on a machine equipped with an AMD Ryzen 9 7940HX CPU and 32 GB of RAM.

In the experiment, SVM and MLKNN were used as classifiers, and evaluation metrics included Micro-F_1, Macro-F_1, Hamming Loss (HL) and Zero-One Loss (ZOL). Specifically, higher values of Micro-F_1 and Macro-F_1 demonstrate improved effectiveness of the corresponding approach, while lower values of HL and ZOL signify better performance of the corresponding method.

5.2. Evaluation metrics

We assess the approach introduced in this study using four widely adopted multi-label metrics: two example-based measures(Hamming Loss and Zero-One Loss), and two label-based measures(Micro-F_1 and Macro-F_1). Lower values of Hamming Loss and Zero-One Loss indicate superior performance, whereas higher values of Micro-F_1 and Macro-F_1 denote better classification quality.

(1) Hamming Loss (HL) functions to measure the discrepancy between the predicted label set and the true label set by calculating the proportion of misclassified labels, thereby evaluating the discriminative effectiveness of the selected features.

$$HL = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y'_{i.} \oplus Y_{i.}\right|}{l}$$

(2) Zero-One Loss measures whether the predicted label set matches the ground truth labels precisely, assigning a loss of 0 for perfect alignment and 1 otherwise, making it suitable for scenarios requiring strict all-or-nothing evaluation.

$$ZOL = \frac{1}{n}\sum_{i=1}^{n}\delta\left(\arg\max_{Y_{.j}\in Y}\mathcal{J}\left(X_{i.};Y_{.j}\right)\right)$$

(3) Micro-F_1 is used to comprehensively evaluate the overall classification performance of the model in all categories by calculating the harmonic mean of precision and recall through aggregated statistics from all classes, making it particularly suitable for global performance evaluation in class-imbalanced scenarios.

$$\text{Micro-}F_1 = \frac{\sum_{i=1}^{l} 2TP^i}{\sum_{i=1}^{l}\left(2TP^i + FP^i + FN^i\right)}$$

(4) Macro-F_1 serves to equally evaluate the classification performance of each category by individually calculating the F_1 score for each class and then taking the arithmetic mean, thereby avoiding dominance by majority classes and making it suitable for scenarios where class importance is balanced.

$$\text{Macro-}F_1 = \frac{1}{l}\sum_{i=1}^{l}\frac{2TP^i}{2TP^i + FP^i + FN^i}$$

In these above evaluation metrics, let Y denote the real label set containing l labels, and let Y' denote the label set predicted through the trained multi-label classifier. $\mathcal{J}(X_{i.};Y_{.j})$ represents the probability that label $Y_{.j}$ is appropriate for instance $X_{i.}$. Additionally, $\oplus$ denotes the XOR operation, and δ is a control flag defined when $\arg\max_{Y_{.j}\in Y}\mathcal{J}(X_{i.};Y_{.j}) \notin Y_i$. Here, δ equals 1 when the condition is met and 0 otherwise. For the i-th label, P, N, T, and F stand for Positive, Negative, True, and False.

5.3. Experimental results

This subsection presents an analysis of all experimental results. The experimental results are shown in Table 3, where bold fonts demonstrate the best results of the proposed method and underlined fonts demonstrate the second-best results of the proposed method on the used datasets. To further evaluate the performance of NCMFS, ten datasets, namely “Bird”, “Business”, “Education”, “Emotions”, “Enron”, “Flags”, “Health”, “Recreation”, “Scene” and “Science”, are selected for comparative analysis under four different evaluation criteria. The specific results are shown in Fig. 2. The following results are observed through the experiment:

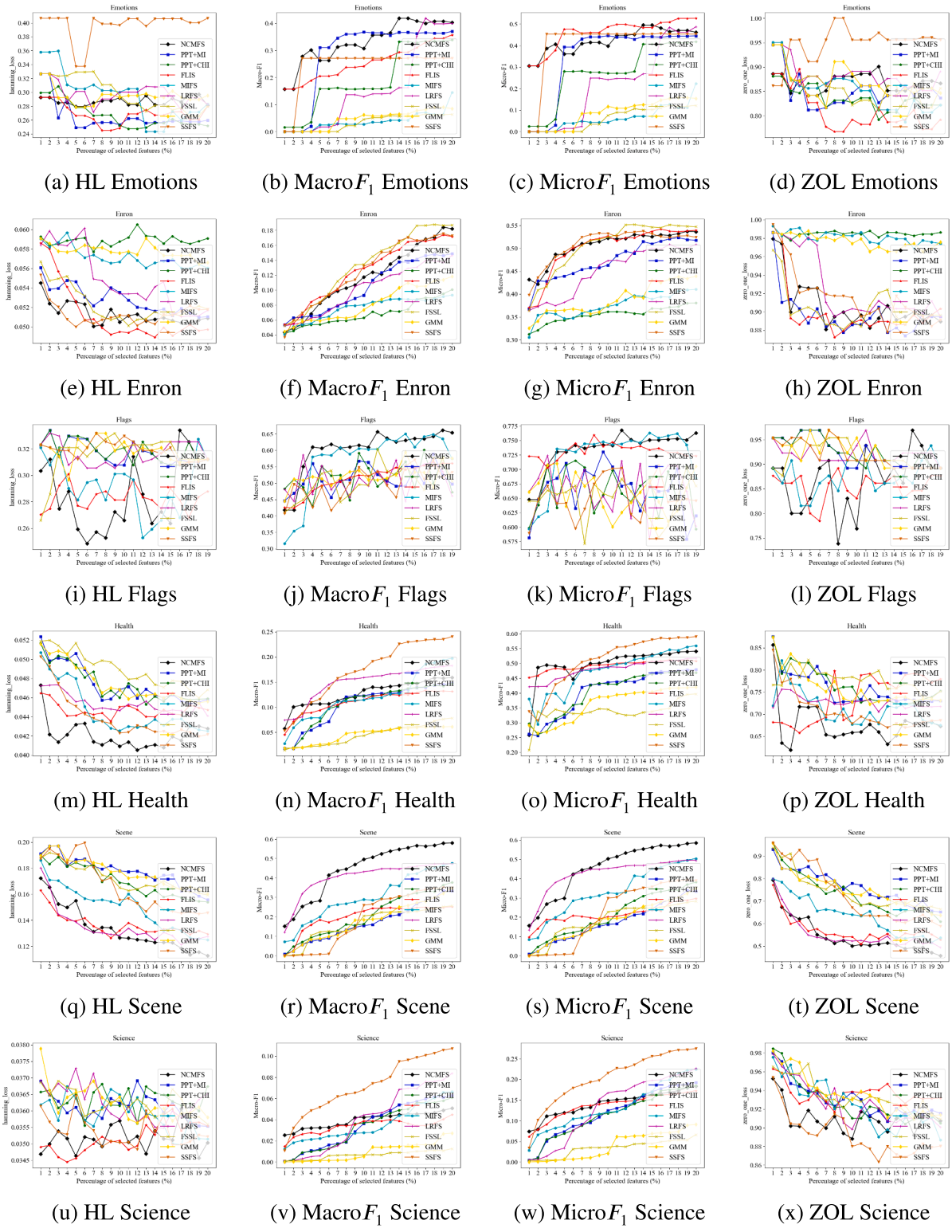
- In Table 3, when using Micro-F_1 as the evaluation metric, NCMFS achieves the optimal and the suboptimal results on the “Emotions”, “Enron”, “Flags”, “Health”, “Scene”, and “Science” datasets. MIFS performs best on the “Business” and “Recreation” datasets. SSFS performs best on the “Education”, “Enron”, “Health”, and “Science” datasets. When Macro-F_1 is used as the evaluation metric, NCMFS is optimal on the “Emotions”, “Flags” and “Scene” datasets, and its performance on dataset “Science” is second only to SSFS. MIFS performs best on the “Business”, and “Recreation” datasets. SSFS has the best result on the “Health” and “Science” data set. FLIS attains the highest performance on the “Birds” and “Education” datasets. FSSL attains the top performance on the “Enron” dataset. For the HL metric, NCMFS shows the best performance on “Birds”, “Enron”, “Flags”, “Health”, “Scene” and “Science” datasets, FLIS are optimal on the “Business” and “Education” datasets. PPT + MI has the best performance on the “Emotions” data set. MIFS has the best performance on the “Recreation” data set. When ZOL as the evaluation metric, SSFS performs best on two datasets, FLIS performs best on three datasets, and NCMFS achieves the optimal and the suboptimal results on the “Business”, “Flags”, “Health”, “Scene” and “Science” datasets, respectively. Through a comprehensive analysis of Table 2 and Table 3, NCMFS demonstrates superior classification performance in the field of image data. However, NCMFS did not achieve optimal results under some evaluation metrics because the fact is that not every feature is correlated with every other feature. Creating a correlation between each pair of features and reconstructing the feature space on that basis is likely to be counterproductive. In some datasets, the classification performance of SSFS is superior to that of NCMFS. This is because SSFS performs Non-negative Matrix Factorization (NMF) on both the feature space and the label space, enables the labels and features to share correlations, and thus greatly enriches the latent feature structure and latent label structure. For the ten binary label datasets used in our study, SSFS achieves better performance on some of them. Although our method also deeply explores the latent feature structure and latent label structure, it is more suitable for scenarios where the labels have ambiguous annotations.

Table 3
Complete experimental results (mean ± std.) showing Micro-averaged F_1-score (Micro-F_1 ↑), Macro-averaged F_1-score(Macro-F_1 ↑), Hamming Loss (HL ↓) and Zero-One Loss (ZOL ↓).

Datasets	NCMFS	MIFS	FLIS	LRFS	GMM	SSFS	PPT + MI	FSSL	PPT + CHI
Micro-$\mathbf{F_1}$ (↑)									
Birds	0.1007 ± 0.0801	0.0000 ± 0.0000	0.0661 ± 0.0391	**0.1727 ± 0.0516**	0.1524 ± 0.0885	0.0735 ± 0.0165	0.0002 ± 0.0017	0.0090 ± 0.0257	0.0064 ± 0.0069
Business	0.6746 ± 0.0024	**0.6835 ± 0.0078**	0.6831 ± 0.0068	0.6796 ± 0.0047	0.6676 ± 0.0000	0.6817 ± 0.0091	0.6729 ± 0.0042	0.6676 ± 0.0000	0.6725 ± 0.0036
Education	0.0721 ± 0.0455	0.0733 ± 0.0587	0.2087 ± 0.0441	0.1495 ± 0.0806	0.0569 ± 0.0415	**0.2500 ± 0.0844**	0.1237 ± 0.0834	0.0656 ± 0.0505	0.1199 ± 0.0848
Emotions	0.4254 ± 0.0504	0.0684 ± 0.0590	**0.4604 ± 0.0671**	0.2257 ± 0.1846	0.0897 ± 0.0608	0.4215 ± 0.1169	0.3442 ± 0.1752	0.0627 ± 0.0514	0.2730 ± 0.1378
Enron	0.5030 ± 0.0373	0.3723 ± 0.0274	0.4957 ± 0.0576	0.4457 ± 0.0555	0.3850 ± 0.0357	**0.5055 ± 0.0417**	0.4695 ± 0.0426	0.5010 ± 0.0631	0.3531 ± 0.0189
Flags	**0.7345 ± 0.0323**	0.7225 ± 0.0499	0.7318 ± 0.0123	0.6659 ± 0.0405	0.6603 ± 0.0321	0.6250 ± 0.0339	0.6646 ± 0.0403	0.6477 ± 0.0329	0.6632 ± 0.0379
Health	0.4924 ± 0.0705	0.4681 ± 0.0795	0.4903 ± 0.0222	0.4754 ± 0.0370	0.3618 ± 0.0522	**0.4999 ± 0.1022**	0.4017 ± 0.0743	0.3347 ± 0.0391	0.3993 ± 0.0748
Recreation	0.0735 ± 0.0316	**0.2524 ± 0.0698**	0.2035 ± 0.0367	0.0392 ± 0.0374	0.0070 ± 0.0108	0.2245 ± 0.0801	0.0343 ± 0.0364	0.0303 ± 0.0179	0.0402 ± 0.0384
Scene	**0.4401 ± 0.1530**	0.3255 ± 0.1346	0.2095 ± 0.0498	0.4146 ± 0.1077	0.1778 ± 0.1021	0.2143 ± 0.1623	0.1677 ± 0.0995	0.1772 ± 0.0954	0.2052 ± 0.1167
Science	0.1384 ± 0.0324	0.1286 ± 0.0569	0.1332 ± 0.0340	0.1286 ± 0.0776	0.0368 ± 0.0320	**0.1985 ± 0.0694**	0.1146 ± 0.0590	0.0320 ± 0.0184	0.1104 ± 0.0546
Macro-$\mathbf{F_1}$ (↑)									
Birds	0.0575 ± 0.0501	0.0000 ± 0.0000	**0.1101 ± 0.0506**	0.0860 ± 0.0267	0.0808 ± 0.0418	0.0565 ± 0.0165	0.0004 ± 0.0028	0.0038 ± 0.0116	0.0038 ± 0.0116
Business	0.0375 ± 0.0045	**0.0564 ± 0.0107**	0.0521 ± 0.0035	0.0547 ± 0.0072	0.0309 ± 0.0000	0.0546 ± 0.0136	0.0385 ± 0.0082	0.0309 ± 0.0000	0.0395 ± 0.0079
Education	0.0124 ± 0.0079	0.0195 ± 0.0170	**0.0657 ± 0.0132**	0.0485 ± 0.0279	0.0132 ± 0.0106	0.0647 ± 0.0202	0.0342 ± 0.0249	0.0205 ± 0.0160	0.0345 ± 0.0263
Emotions	**0.3358 ± 0.0711**	0.0409 ± 0.0365	0.2563 ± 0.0632	0.1660 ± 0.1549	0.0483 ± 0.0329	0.2521 ± 0.0699	0.2825 ± 0.1450	0.0346 ± 0.0277	0.1925 ± 0.1176
Enron	0.1141 ± 0.0437	0.0743 ± 0.0169	0.1224 ± 0.0414	0.0988 ± 0.0325	0.0798 ± 0.0237	0.1224 ± 0.0447	0.1019 ± 0.0346	**0.1293 ± 0.0489**	0.0670 ± 0.0164
Flags	**0.6024 ± 0.0677**	0.5697 ± 0.0998	0.5137 ± 0.0428	0.5106 ± 0.0442	0.5233 ± 0.0329	0.4987 ± 0.0480	0.5135 ± 0.0372	0.5060 ± 0.0511	0.5174 ± 0.0457
Health	0.1237 ± 0.0312	0.1181 ± 0.0461	0.1078 ± 0.0270	0.1454 ± 0.0378	0.0441 ± 0.0164	**0.1685 ± 0.0628**	0.1023 ± 0.0435	0.0431 ± 0.0198	0.1039 ± 0.0428
Recreation	0.0321 ± 0.0132	**0.1336 ± 0.0334**	0.0228 ± 0.0124	0.0231 ± 0.0225	0.0043 ± 0.0065	0.1119 ± 0.0422	0.0197 ± 0.0213	0.0148 ± 0.0088	0.0245 ± 0.0239
Scene	**0.4325 ± 0.1533**	0.2947 ± 0.1268	0.1945 ± 0.0754	0.3942 ± 0.1039	0.1736 ± 0.1002	0.1858 ± 0.1437	0.1634 ± 0.0976	0.1589 ± 0.0868	0.1991 ± 0.1137
Science	0.0389 ± 0.0077	0.0341 ± 0.0164	0.0329 ± 0.0076	0.0388 ± 0.0287	0.0087 ± 0.0081	**0.0695 ± 0.0280**	0.0335 ± 0.0221	0.0064 ± 0.0035	0.0318 ± 0.0205
HL (↓)									
Birds	**0.0516 ± 0.0018**	0.0528 ± 0.0008	0.0520 ± 0.0014	0.0540 ± 0.0005	0.0519 ± 0.0011	0.0577 ± 0.0045	0.0521 ± 0.0017	0.0521 ± 0.0017	0.0536 ± 0.0009
Business	0.0292 ± 0.0003	0.0288 ± 0.0006	**0.0286 ± 0.0004**	0.0288 ± 0.0005	0.0296 ± 0.0005	0.0292 ± 0.0006	0.0296 ± 0.0004	0.0295 ± 0.0005	0.0296 ± 0.0004
Education	0.0449 ± 0.0007	0.0449 ± 0.0007	**0.0427 ± 0.0008**	0.0447 ± 0.0011	0.0450 ± 0.0008	0.0632 ± 0.0019	0.0448 ± 0.0013	0.0457 ± 0.0009	0.0446 ± 0.0014
Emotions	0.2874 ± 0.0053	0.2948 ± 0.0346	0.2700 ± 0.0134	0.2966 ± 0.0151	0.2948 ± 0.0107	0.3990 ± 0.0175	**0.2647 ± 0.0197**	0.2983 ± 0.0236	0.2671 ± 0.0191
Enron	**0.0515 ± 0.0016**	0.0574 ± 0.0012	0.0516 ± 0.0031	0.0550 ± 0.0030	0.0577 ± 0.0007	**0.0515 ± 0.0015**	0.0528 ± 0.0015	0.0525 ± 0.0022	0.0589 ± 0.0005
Flags	**0.2851 ± 0.0261**	0.2920 ± 0.0205	0.2891 ± 0.0103	0.3169 ± 0.0087	0.3181 ± 0.0082	0.3191 ± 0.0065	0.3171 ± 0.0082	0.3157 ± 0.0151	0.3184 ± 0.0083
Health	**0.0422 ± 0.0021**	0.0448 ± 0.0026	0.0448 ± 0.0011	0.0454 ± 0.0013	0.0476 ± 0.0021	0.0440 ± 0.0024	0.0476 ± 0.0021	0.0491 ± 0.0018	0.0476 ± 0.0020
Recreation	0.0645 ± 0.0008	**0.0601 ± 0.0017**	0.0663 ± 0.0011	0.0667 ± 0.0007	0.0670 ± 0.0001	0.0630 ± 0.0011	0.0667 ± 0.0006	0.0656 ± 0.0006	0.0670 ± 0.0008
Scene	**0.1350 ± 0.0192**	0.1507 ± 0.0173	0.1382 ± 0.0105	0.1362 ± 0.0160	0.1771 ± 0.0106	0.1658 ± 0.0192	0.1780 ± 0.0105	0.1734 ± 0.0105	0.1708 ± 0.0126
Science	**0.0352 ± 0.0003**	0.0360 ± 0.0006	0.0352 ± 0.0004	0.0362 ± 0.0005	0.0363 ± 0.0004	0.0354 ± 0.0004	0.0362 ± 0.0004	0.0362 ± 0.0004	0.0364 ± 0.0003
ZOL (↓)									
Birds	0.5352 ± 0.0189	0.5480 ± 0.0103	0.5341 ± 0.0152	**0.5270 ± 0.0031**	0.5335 ± 0.0108	0.5759 ± 0.0333	0.5475 ± 0.0183	0.5485 ± 0.0178	0.5533 ± 0.0069
Business	0.4810 ± 0.0077	0.4821 ± 0.0119	**0.4734 ± 0.0104**	0.4847 ± 0.0153	0.4896 ± 0.0130	0.4894 ± 0.0111	0.4916 ± 0.0112	0.4836 ± 0.0130	0.4951 ± 0.0101
Education	0.9354 ± 0.0275	0.9421 ± 0.0373	0.9204 ± 0.0228	0.9105 ± 0.0304	0.9283 ± 0.0279	**0.8635 ± 0.0317**	0.9162 ± 0.0287	0.9224 ± 0.0335	0.9164 ± 0.0351
Emotions	0.8678 ± 0.0172	0.8487 ± 0.0398	**0.8112 ± 0.0399**	0.8819 ± 0.0274	0.8656 ± 0.0310	0.9501 ± 0.0301	0.8494 ± 0.0328	0.8550 ± 0.0339	0.8335 ± 0.0222
Enron	0.9082 ± 0.0305	0.9817 ± 0.0064	**0.9007 ± 0.0343**	0.9264 ± 0.0368	0.9765 ± 0.0078	0.9178 ± 0.0306	0.9011 ± 0.0316	0.9055 ± 0.0266	0.9847 ± 0.0027
Flags	**0.8688 ± 0.0587**	0.8761 ± 0.0321	0.8698 ± 0.0253	0.9206 ± 0.0182	0.9223 ± 0.0280	0.9190 ± 0.0263	0.9223 ± 0.0293	0.9166 ± 0.0251	0.9255 ± 0.0275
Health	**0.6768 ± 0.0540**	0.7108 ± 0.0329	0.7294 ± 0.0486	0.7278 ± 0.0285	0.7611 ± 0.0385	0.7103 ± 0.0511	0.7572 ± 0.0345	0.7782 ± 0.0285	0.7586 ± 0.0389
Recreation	0.9380 ± 0.0262	**0.8223 ± 0.0459**	0.9553 ± 0.0220	0.9555 ± 0.0188	0.9612 ± 0.0146	0.8656 ± 0.0405	0.9566 ± 0.0192	0.9400 ± 0.0259	0.9552 ± 0.0202
Scene	**0.5562 ± 0.1008**	0.6390 ± 0.0982	0.5824 ± 0.0790	0.5615 ± 0.0903	0.7597 ± 0.0795	0.7404 ± 0.1395	0.7696 ± 0.0809	0.7512 ± 0.0920	0.7345 ± 0.0934
Science	0.9132 ± 0.0191	0.9263 ± 0.0271	0.9362 ± 0.0142	0.9288 ± 0.0260	0.9361 ± 0.0234	**0.8961 ± 0.0286**	0.9316 ± 0.0215	0.9356 ± 0.0196	0.9289 ± 0.0242

- In Fig. 2, on the X-axis is the number of selected features, and on the Y-axis is the classification performance measured by the evaluation metrics using different methods. Fig. 2(a) to Fig. 2(d) show line charts of the "Emotions" dataset across four evaluation metrics. In NCMFS, the Macro-F_1 and Micro-F_1 values of the "Emotions" dataset outperform those of other methods. Fig. 2(e) to Fig. 2(h) present line charts of the "Enron" dataset across four evaluation metrics. For the "Enron" dataset, Micro-F_1 and ZOL outperform other methods. Fig. 2(i) to Fig. 2(l) display line charts of the "Flags" dataset across four evaluation metrics. In NCMFS, the HL, Macro-F_1, Micro-F_1, and ZOL of the "Flags" dataset exceed other methods. Fig. 2(m) to Fig. 2(p) are line charts of the "Health" dataset in four evaluation metrics. In NCMFS, the Micro-F_1, HL and ZOL values of the "Health" dataset all outperform other methods. Fig. 2(q) to Fig. 2(t) show line charts of the "Scene" dataset across four evaluation metrics. In NCMFS, the HL, Macro-F_1, Micro-F_1, and ZOL values of the "Scene" dataset outperform other methods. Fig. 2(u) to Fig. 2(x) present line charts of the "Science" dataset in four evaluation metrics. In NCMFS, although the Macro-F_1, Micro-F_1 and ZOL is inferior to SSFS, the performance of NCMFS is good compared to other methods. Based on the above analysis, NCMFS demonstrates superiority in the field of image classification. However, NCMFS did not achieve optimal results under some evaluation metrics because the fact is that not every feature is correlated with every other feature. Creating a correlation between each pair of features and reconstructing the feature space on that basis is likely to be counterproductive.
- The superior performance of NCMFS demonstrates that the negative label space information and the sparsity regularization term can capture the intricate dependencies between features and labels more comprehensively and accurately. Meanwhile, NCMFS construct a feature correlation matrix to characterize the inter-dependencies among features and thereby reconstruct the feature space. The subsequent ablation study confirms these findings.

We further analyze the relative performance of NCMFS against competing methods using the Friedman test. First, all average ranks are summarized in Table 4. Next, the Friedman statistics F_F and the corresponding critical values for each evaluation metric are reported



(a) HL Emotions (b) MacroF_1 Emotions (c) MicroF_1 Emotions (d) ZOL Emotions

(e) HL Enron (f) MacroF_1 Enron (g) MicroF_1 Enron (h) ZOL Enron

(i) HL Flags (j) MacroF_1 Flags (k) MicroF_1 Flags (l) ZOL Flags

(m) HL Health (n) MacroF_1 Health (o) MicroF_1 Health (p) ZOL Health

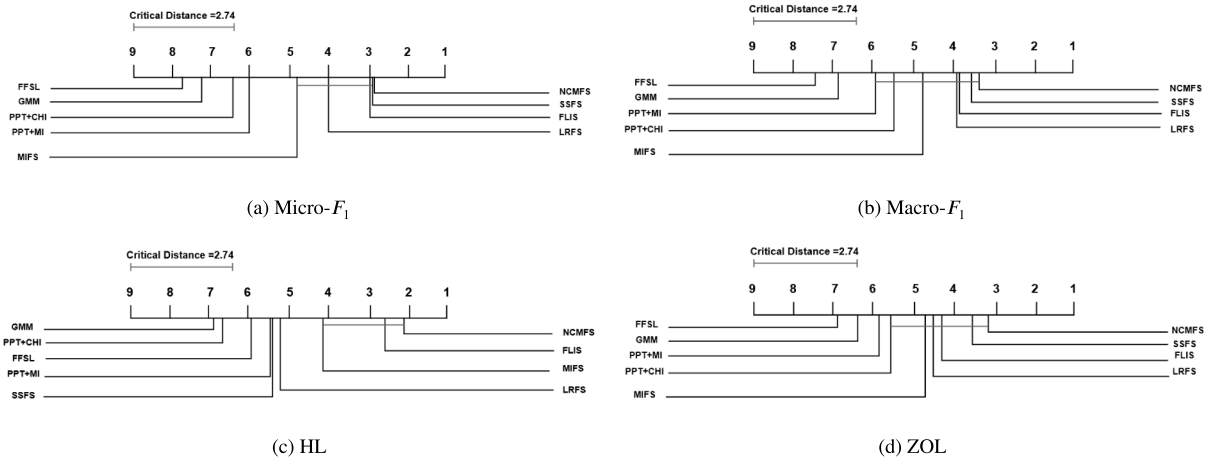
(q) HL Scene (r) MacroF_1 Scene (s) MicroF_1 Scene (t) ZOL Scene

(u) HL Science (v) MacroF_1 Science (w) MicroF_1 Science (x) ZOL Science

Fig. 2. The experimental results in terms of Micro-F_1, Macro-F_1, HL and ZOL.

in Table 5. Based on Table 5, at the significance level $\alpha = 0.1$, the null hypothesis is rejected, which means that all methods perform not equally.

Because our goal is to assess the pairwise relationship between NCMFS (as the control method) and the other methods, we adopt the Bonferroni–Dunn test as a post-hoc procedure. By examining whether the difference between two methods' average ranks falls within the Critical Distance (CD), we determine whether NCMFS differs significantly from the others. The CD is computed as $\mathrm{CD} = q_\alpha \sqrt{k(k+1)/(6N)}$ with $q_\alpha = 2.24$, $k = 9$ (methods), and $N = 10$ (datasets), which gives $\mathrm{CD} = 2.74$. Fig. 3 presents the CD diagrams for all evaluation metrics. The red line indicates the range of one CD from NCMFS, reflecting statistical similarity within that interval. As can be observed, although NCMFS, SSFS, and FLIS do not differ significantly from one another,



(a) Micro-F_1 (b) Macro-F_1

(c) HL (d) ZOL

Fig. 3. Average rank graph form of Bonferroni-Dunn test results.

Table 4
The average ranks of nine methods in terms of each evaluation criterion.

Metric	NCMFS	MIFS	FLIS	LRFS	GMM	SSFS	PPT + MI	FSSL	PPT + CHI
Micro-F_1	2.9	4.8	3	4	7.2	2.9	6	7.6	6.4
Macro-F_1	3.5	4.7	3.8	3.8	6.8	3.6	5.9	7.3	5.4
HL	2.2	4.1	2.5	5.2	6.9	5.4	5.4	5.9	6.7
ZOL	3.3	4.3	3.6	4.5	6.9	4.6	5.8	5.6	6.3

Table 5
Summary of Friedman statistics and critical values for each evaluation metric.

Metric	$\mathbf{F_F}$	Critical value ($\boldsymbol{\alpha = 0.1}$)
Micro-F_1(SVM)	11.015	1.757
Macro-F_1(SVM)	4.535	
HL	4.906	
ZOL	3.160	

NCMFS shows a competitive advantage over the remaining methods; moreover, the proposed NCMFS ranks first across all evaluation metrics.

5.4. Ablation study

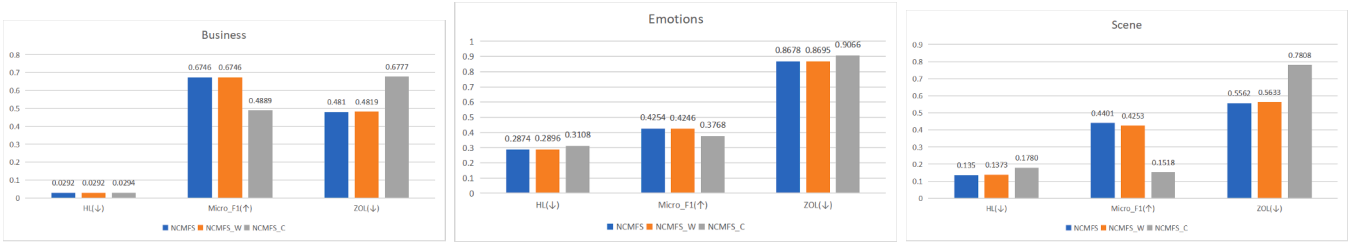
To validate the contribution of each NCMFS component, we performed the ablation study. As shown in Fig. 4, NCMFS is the proposed method. NCMFS$_W$ denotes the sparsity regularization term after ablating $\hat{W}$ in Formula 6. NCMFS$_C$ denotes the formula after ablating the feature correlation matrix in Formula 6. Three datasets, namely "Business", "Emotions" and "Scene", are selected for ablation study under three different evaluation criteria.

Impact of the reconstructed feature space. The ablation study demonstrates that NCMFS beats NCMFS$_C$ in most cases. The reconstructed feature space contains richer information than the original.

Impact of the sparsity regularization term based on the difference between the original weight matrix and the mirrored weight matrix . The ablation results demonstrate that although the results of NCMFS and NCMFS$_W$ are quite similar, NCMFS consistently shows a slight advantage over NCMFS$_W$. Since the X matrix is fixed and considering the terms XW and $X\hat{W}$, therefore, the sparsity regularization term based on the difference between the original weight matrix and the mirrored weight matrix eliminates the impact of ambiguous labels. Since all ten datasets used in the study are hard-label datasets, ambiguous labels can only be approximated by XW. Therefore, the effect of the method in eliminating ambiguous features is not significant. If the weight values of certain ambiguous labels are between 0 and 1, the sparsity regularization term based on the difference between the original weight matrix and the mirrored weight matrix alleviates the interference of ambiguous labels since the larger the difference between W and $\hat{W}$, the greater the distinction between the corresponding labels.

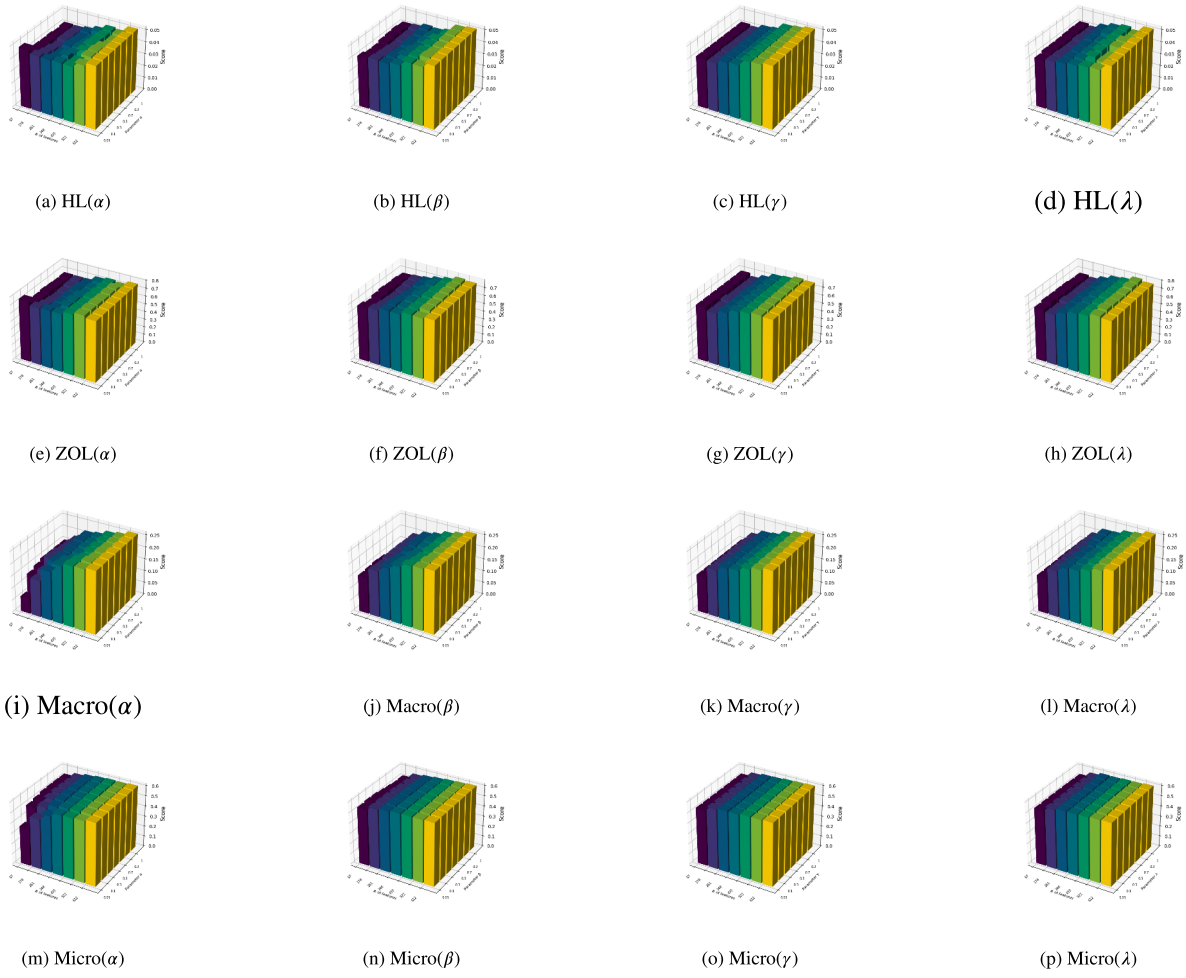
5.5. Parameter sensitivity analysis

In NCMFS, the Health dataset is selected to conduct a parameter



(a) Business (b) Emotions (c) Scene

Fig. 4. Ablation experimental results.



(a) HL(α) (b) HL(β) (c) HL(γ) (d) HL(λ)

(e) ZOL(α) (f) ZOL(β) (g) ZOL(γ) (h) ZOL(λ)

(i) Macro(α) (j) Macro(β) (k) Macro(γ) (l) Macro(λ)

(m) Micro(α) (n) Micro(β) (o) Micro(γ) (p) Micro(λ)

Fig. 5. Parameter Sensitivity Analysis on the Health.

sensitivity experiment by setting four regularization parameters: α, β, γ and λ. In the experiment, only one parameter is adjusted each time, with the adjustment range being the grid $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$, and the remaining parameters are all fixed at 0.5. The experimental results are shown in Fig. 5. Among them, Fig. 5(a)-(h) present the parameter sensitivity results of the MLKNN classifier under the Hamming Loss (HL) and Zero-One Loss (ZOL) metrics; Fig. 5(i)-(p) show the results of the SVM classifier under the Micro-F_1 and Macro-F_1 metrics. Observing the changes in the histograms in the figures, it is apparent that NCMFS is insensitive to the three parameters β, γ and λ, but sensitive to the parameter α. The reason why NCMFS is sensitive to the parameterα: (1) α controls how close C must be to an identity mapping through $\|\mathbf{XC} - \mathbf{X}\|_F^2$. Since XC also appears inside the supervised terms $\left\|\mathbf{XC}\hat{\mathbf{W}} - \hat{\mathbf{Y}}\right\|_F^2$ and $\left\|\mathbf{I}_{n\times l} - \mathbf{XCW} - \mathbf{XC}\hat{\mathbf{W}}\right\|_F^2$, changing α re-scales the effective strength of those terms indirectly. (2) In the multiplicative updates, the C-update contains $\alpha\, X^T XC$. If $X^T X$ is ill-conditioned, small changes in α amplify step sizes, making the objective highly sensitive to α.

5.6. Convergence and time complexity analysis

We conduct experiments on six benchmark datasets (Emotions, Enron, Entertain, Health, Scene and Science) to analyze the convergence of NCMFS, MIFS, and SSFS in Fig. 6. From the experimental results in the figure, it can be observed that compared with MIFS, NCMFS can converge quickly with only a small number of iterations. However, compared with SSFS, NCMFS has a slightly longer convergence time. The reason for this is that the update of SSFS's objective function is relatively simple, while our method, although its update process is relatively complex, can better explore the connections between features and labels and mitigate the impact of ambiguous labels. The NCMFS values are seen to decline rapidly in the initial iterations and subsequently approach convergence at a slower pace. This trend confirms the effectiveness of the optimization approach outlined in Section 4.

Table 6
Time complexity of methods.

Methods	Negative label information
NCMFS	$\mathcal{O}(d^3 + d^2 n)$
MIFS	$\mathcal{O}(knd + n^2)$
FLIS	$\mathcal{O}(nd + nd^2)$
LRFS	$\mathcal{O}(dl^2 + hd)$
FSSL	$\mathcal{O}(hndl)$
GMM	$\mathcal{O}(ndl)$
SSFS	$\mathcal{O}(ndh + n^2 h)$
PPT + MI	$\mathcal{O}(nd)$
PPT + CHI	$\mathcal{O}(nd)$

It is assumed that the dataset contains n samples, with d features, l labels, and the number of selected features denoted as h; k represents the number of clusters in the MIFS method. As can be seen from the Table 6, the time complexity of information-theoretic methods is lower than that of embedded methods, because embedded methods need to consider more high-order correlations. Embedded methods involve

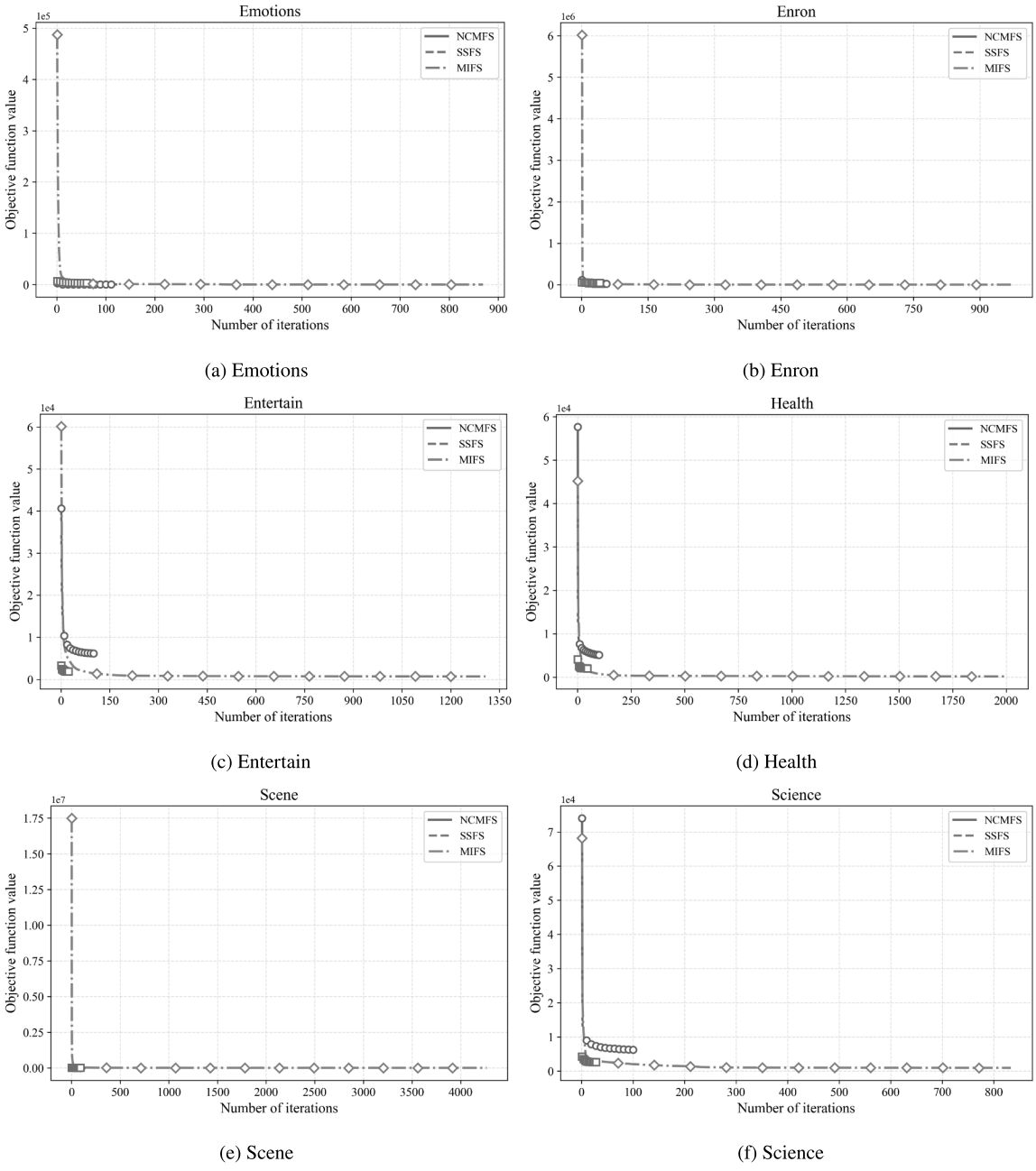


Fig. 6. Convergence analysis.

Table 7
The running time (s) of all methods.

Datasets	NCMFS	MIFS	FLIS	LRFS	GMM	SSFS	PPT + MI	FSSL	PPT + CHI
Birds	0.7274	4.3264	370.0569	153.8198	1.9339	0.3600	0.0738	16.7018	2.554
Business	7.4259	27.9942	9320.2408	4473.9806	32.1042	2.8484	1.0949	764.2858	65.1740
Education	14.8027	35.6676	15753.9513	7306.1785	43.2493	6.7599	1.2427	1869.3865	93.0931
Emotions	0.1903	5.4764	10.7044	4.4760	0.2124	0.2483	0.0319	0.6911	1.0023
Enron	329.3658	155.9359	45250.8533	18092.3621	70.1554	6.5934	0.7879	1210.5178	13.1508
Flags	0.1853	37.7202	0.9694	0.5316	0.0309	0.1855	0.0019	0.3759	0.2075
Health	15.4064	110.7948	19364.8619	8433.4837	47.0652	5.8503	1.3803	2294.4421	93.0911
Recreation	15.1497	29.0872	11941.1351	5213.5564	32.5140	5.6529	1.3674	1332.2774	73.9912
Scene	2.6219	27.0767	476.1407	346.0897	4.7493	3.7250	0.7799	323.9220	659.0871
Science	21.1504	20.2099	34527.7357	16294.1888	73.5574	5.9112	1.9797	3129.1967	196.2386

fewer iterations and thus have shorter actual running time. Although the time complexity of NCMFS per iteration is not the lowest, its classification performance is superior to that of other compared methods. Additionally, it requires fewer iterations and has an advantage in convergence efficiency.

Table 7 presents the running time (in seconds) of each comparison method on all used datasets, which is the total time recorded from the start to the end of each method after the training data is imported into it. It can be seen that the SSFS and PPT + MI methods have the shortest running time because they do not focus on label correlations; however, the classification performance of NCMFS is superior to these two methods. The running time of NCMFS is shorter than that of MIFS and GMM, and both of these methods belong to embedded methods. The remaining methods are all mutual information-based feature selection methods with longer running time. Overall, NCMFS is superior to all comparison methods in both effectiveness and efficiency.

6. Conclusion

In this paper, we introduced Negative Label-Aware and Correlation-Enhanced Multi-Label Feature Selection (NCMFS), a novel framework that uses negative label information to improve the selection of informative features in multi-label settings. By simultaneously using feature space information to align the primary label space with the mirrored label space, NCMFS captures complex feature-label relationships more comprehensively and enhances feature discriminability. To further refine the selection process, we incorporated sparsity regularization on the difference between the original weight matrix and its mirrored counterpart. This strategy effectively mitigates the interference caused by ambiguous labels. Finally, we construct a feature correlation matrix to characterize the inter-dependencies among features and thereby reconstruct the feature space, ensuring strongly correlated features are emphasized while weakly correlated features are downplayed. Extensive experiments on benchmark datasets demonstrate that NCMFS consistently outperforms existing methods with respect to precision, sparsity, and calculative efficiency. These results validate the effectiveness of mining the negative label information and the informative feature space.

Despite NCMFS's strong results, redundant features in multi-label data remain a key barrier to further boosting downstream classification performance. Therefore, we aim to integrate domain-specific knowledge to further optimize performance.

CRediT authorship contribution statement

Xiang Li: Writing – review & editing, Writing – original draft, Methodology; **Huimin Fu:** Writing – review & editing, Writing – original draft, Investigation; **Xiaoou Huang:** Validation, Software; **Tianyi Xie:** Validation, Software; **Lingfei Ren:** Formal analysis; **Wanfu Gao:** Supervision, Resources; **Yonghao Li:** Supervision, Conceptualization; **Xin Yang:** Supervision, Resources.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62206227 and 62476228, in part by the Sichuan Science and Technology Program under Grants 2025ZNSFSC1497, 2022NSFSC0528, and 2022ZYD0113, and in part by the Fundamental Research Funds for the Central Universities, Jilin University, under Grant 93K172025K12.

References

[1] B. Jiang, J. Liu, Z. Wang, C. Zhang, J. Yang, Y. Wang, W. Sheng, W. Ding, Semi-supervised multi-view feature selection with adaptive similarity fusion and learning, Pattern Recognit. 159 (2025) 111159.
[2] Y. Zhang, J. Tang, Z. Cao, Semi-supervised multi-label feature selection via partial label correlation and feature self-representation, Knowl. Based Syst. (2025) 113632.
[3] Z. Sun, Z. Chen, J. Liu, Y. Chen, Y. Yu, Partial multi-label feature selection via low-rank and sparse factorization with manifold learning, Knowl. Based Syst. 296 (2024) 111899.
[4] X. Gong, J. Wang, Q. Ren, K. Zhang, E.-S.M. El-Alfy, J. Mańdziuk, Embedded feature selection approach based on TSK fuzzy system with sparse rule base for high-dimensional classification problems, Knowl. Based Syst. 295 (2024) 111809.
[5] Z. He, Y. Lin, Z. Lin, C. Wang, Multi-label feature selection via similarity constraints with non-negative matrix factorization, Knowl. Based Syst. 297 (2024) 111948.
[6] R. Sheikhpour, M.A. Sarram, S. Gharaghani, M.A.Z. Chahooki, A survey on semi-supervised feature selection methods, Pattern Recognit. 64 (2017) 141–158.
[7] H. Dong, J. Sun, T. Li, R. Ding, X. Sun, A multi-objective algorithm for multi-label filter feature selection problem, Appl. Intell. 50 (2020) 3748–3774.
[8] J. González, J. Ortega, M. Damas, P. Martín-Smith, J.Q. Gan, A new multi-objective wrapper method for feature selection–accuracy and stability analysis for BCI, Neurocomputing 333 (2019) 407–418.
[9] R.B. Pereira, A. Plastino, B. Zadrozny, L.H.C. Merschmann, Categorizing feature selection methods for multi-label classification, Artif. Intell. Rev. 49 (2018) 57–78.
[10] R. Huang, Z. Wu, Multi-label feature selection via manifold regularization and dependence maximization, Pattern Recognit. 120 (2021) 108149.
[11] J. Gonzalez-Lopez, S. Ventura, A. Cano, Distributed multi-label feature selection using individual mutual information measures, Knowl. Based Syst. 188 (2020) 105052.
[12] J. Liu, Y. Li, W. Weng, J. Zhang, B. Chen, S. Wu, Feature selection for multi-label learning with streaming label, Neurocomputing 387 (2020) 268–278.
[13] J. Zhong, R. Shang, F. Zhao, W. Zhang, S. Xu, Negative label and noise information guided disambiguation for partial multi-Label learning, IEEE Trans. Multimedia 26 (2024) 9920–9935.
[14] W. Gao, Y. Li, L. Hu, Multilabel feature selection with constrained latent structure shared term, IEEE Trans. Neural Netw. Learn. Syst. 34 (3) (2023) 1253–1262.
[15] P. Zhang, G. Liu, W. Gao, Distinguishing two types of labels for multi-label feature selection, Pattern Recognit. 95 (2019) 72–82.
[16] J. Dai, W. Chen, Y. Qian, Multi-Label feature selection with missing features via implicit label replenishment and positive correlation feature recovery, IEEE Trans. Knowl. Data Eng. 37 (4) (2025) 2042–2055.
[17] Y. Lin, Q. Hu, J. Liu, J. Duan, Multi-label feature selection based on max-dependency and min-redundancy, Neurocomputing 168 (2015) 92–103.
[18] L. Hu, Y. Li, W. Gao, P. Zhang, J. Hu, Multi-label feature selection with shared common mode, Pattern Recognit. 104 (2020) 107344.
[19] Y. Li, L. Hu, W. Gao, Multi-label feature selection via robust flexible sparse regularization, Pattern Recognit. 134 (2023) 109074.
[20] G. Doquire, M. Verleysen, Feature selection for multi-label classification problems, in: International Work-conference on Artificial Neural Networks, Springer, 2011, pp. 9–16.
[21] J. Read, A pruned problem transformation method for multi-label classification, in: Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008), 143150, 2008, p. 41.
[22] Y. Lin, Q. Hu, J. Liu, J. Li, X. Wu, Streaming feature selection for multilabel learning based on fuzzy mutual information, IEEE Trans. Fuzzy Syst. 25 (6) (2017) 1491–1507.
[23] L. Jian, J. Li, K. Shu, H. Liu, Multi-label informed feature selection, in: Proceedings of the 25Th International Joint Conference on Artificial Intelligence (IJCAI), 16, 2016, pp. 1627–1633.
[24] Z. Zhang, Z. Zhang, J. Yao, L. Liu, J. Li, G. Wu, X. Wu, Multi-label feature selection via adaptive label correlation estimation, ACM Trans. Knowl. Discov. Data 17 (9) (2023) 1–28.
[25] Z. Cai, W. Zhu, Multi-label feature selection via feature manifold learning and sparsity regularization, Int. J. Mach. Learn. Cybern. 9 (2018) 1321–1334.
[26] Y. Fan, B. Chen, W. Huang, J. Liu, W. Weng, W. Lan, Multi-label feature selection based on label correlations and feature redundancy, Knowl. Based Syst. 241 (2022) 108256.
[27] Y. Fan, J. Liu, J. Tang, P. Liu, Y. Lin, Y. Du, Learning correlation information for multi-label feature selection, Pattern Recognit. 145 (2024) 109899.
[28] D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Adv. Neural Inf. Process. Syst. 13 (2000).
[29] C.H.Q. Ding, T. Li, M.I. Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2008) 45–55.
[30] S. Zhang, Y. Li, P. Zhang, W. Gao, Exploring multi-label feature selection via feature and label information supplementation, Eng. Appl. Artif. Intell. 159 (2025) 111552.