Contents lists available at ScienceDirect

Expert Systems With Applications

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

Multi-label feature selection with global and local label correlation

Mohammad Faraji, Seyed Amjad Seyedi, Fardin Akhlaghian Tab [*], Reza Mahmoodi

Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran

ARTICLE INFO

Keywords:
Feature selection
Multi-label learning
Label correlation
Nonnegative matrix factorization

ABSTRACT

In various application domains, high-dimensional multi-label data has become more prevalent, presenting two significant challenges: instances with high-dimensional features and a large number of labels. In the context of multi-label feature selection, the objective is to choose a subset of features from a given set that is highly pertinent for predicting multiple labels or categories associated with each instance. However, certain characteristics of multi-label classification, such as label dependencies and imbalanced label distribution, have often been overlooked although they hold valuable insights for designing effective multi-label feature selection algorithms. In this paper, we propose a feature selection model which exploits explicit global and local label correlations to select discriminative features across multiple labels. In addition, by representing the feature matrix and label matrix in a shared latent space, the model aims to capture the underlying correlations between features and labels. The shared representation can reveal common patterns or relationships that exist across multiple labels and features. An objective function involving $L_{2,1}$-norm regularization is formulated, and an alternating optimization-based iterative algorithm is designed to obtain the sparse coefficients for multi-label feature selection. The proposed method was evaluated on 14 real-world multi-label datasets using six evaluation metrics, through comprehensive experiments. The results indicate its effectiveness, surpassing that of several representative methods.

1. Introduction

In recent years, feature selection has become increasingly important due to the detrimental effects of the high dimensionality curse. These effects include elevated learning complexity (Li, Miao, & Pedrycz, 2017; Liu, Lin, Li, Weng and Wu, 2018), heightened space allocation (Lin, Hu, Liu, Li, & Wu, 2017), and reduced classification performance (Gao, Hu, & Zhang, 2018). The number of features can be decreased and classification accuracy can be increased by choosing only pertinent features and eliminating redundant and irrelevant ones (Huang, Li, Huang, & Wu, 2017; Zhu, Xu, Hu, Zhang, & Zhao, 2018). The use of feature selection algorithms has been widespread in several disciplines, including recommender system (Abdollahi, Amjad Seyedi, & Reza Noorimehr, 2020; Shajarian, Seyedi, & Moradi, 2017; Tang, Kay, & He, 2016), picture and video annotation (Hong et al., 2013), microarray data processing, and genomics (Xie, Wang, Xu, Huang, & Grant, 2021). Since objects in the real world can have numerous labels at once, substantial research has been done on feature selection in these domains. For instance, a single gene in genomics can perform several tasks, including photosynthesis, protein breakdown, and signal transmission (Wang & Domeniconi, 2008). In music analysis, one piece of music can have several different emotions at the same time, like sadness, joy, and scariness (Trohidis, Tsoumakas, Kalliris, & Vlahavas, 2011), and how a newswire story can be classified in news categorization.

Multi-label learning typically makes the correlation between the labels an assumption, unlike multi-class problems. Because of this, extracting and using the correlation of labels, which has made it into an NP-hard problem, is the main challenge in multi-label learning in addition to the dimension of label space. Currently used techniques for selecting multi-label features emphasize the extraction of label correlation, label-feature relevance, and feature correlation. Existing approaches for multi-label feature selection are classified into two groups in order to handle multi-label data: problem transformation models and algorithm adaption models (Zhang & Zhou, 2014). In problem transformation, multi-label data is converted to single-label data using problem transformation techniques, then single-label feature selection methods are applied to the single-label data. To directly handle multi-label data, researchers have also offered a number of algorithm adaption models for multi-label feature selection (Jian, Li, Shu, & Liu, 2016; Zhang & Ma, 2022). Label correlations are primarily extracted by algorithm adaptation techniques to direct the feature selection procedure. Adaptation methods outperform problem transformation methods that ignore label correlations in terms of classification performance because they take label correlation into account. The three approaches

[*] Corresponding author.
E-mail addresses: mohammad.faraji@uok.ac.ir (M. Faraji), amjadseyedi@uok.ac.ir (S.A. Seyedi), f.akhlaghian@uok.ac.ir (F. Akhlaghian Tab), reza.mahmoodi@uok.ac.ir (R. Mahmoodi).

https://doi.org/10.1016/j.eswa.2024.123198
Received 17 September 2023; Received in revised form 9 December 2023; Accepted 7 January 2024

for algorithm adoption are first-order, second-order, and high-order. The correlation between labels is not taken into account in the first-order approach because labels are thought of as independent of one another. The second-order approach takes into account the correlation between label pairs, and in practice, we will have a ranking between labels that are related and those that are not. The correlation between a subset or the entire set of labels is taken into account in the high-order approach.

In multi-label learning, the presence of label correlations can yield significant insights. For instance, if the labels "Ski" and "Snow" exist, it is highly likely that the label "Winter Sport" will also be there. Likewise, if the labels "hot" and "sunny" are both present, it is highly unlikely that the label "snow" will be appear. Incorporating label correlations of various degrees is a goal of recent studies on multi-label learning (Zhou, 2018). Some studies (Fürnkranz, Hüllermeier, Loza Mencía, & Brinker, 2008; Ji, Tang, Yu, & Ye, 2008; Read, Pfahringer, Holmes, & Frank, 2011) concentrate mainly on global label correlation that apply to all instances. However, some label correlation (Huang & Zhou, 2012; Weng, Lin, Wu, Li, & Kang, 2018) are unique to a local data subset. For instance, in internet scope, the word "Amazon" refers to "Amazon shop", whereas in Nature scope, the word "Amazon" refers to "Amazon Forest". Global or local label correlations have been the focus of prior research. But it is clearly better and more desirable to take into account both of them in the Multi-Label Feature Selection problem.

In light of the above analysis, this paper presents a novel feature selection method named Multi-Label Feature Selection with Global and Local label correlation (MLFS-GLOCAL). In this method, the label information is mapped into a low-dimensional reduced space that captures the implicit correlations among multiple labels. In the light of this assumption that correlated features must share similar labels, we map feature information into the same low-dimensional reduced space. Therefore, this method embeds the label and feature correlations into a shared space. Though this low-rank structure can be regarded as implicitly exploiting label correlations, there is still a potential capacity to consider label correlations explicitly. Therefore, we extract both global and local label correlations from training label information and it encourages the prediction to be similar on highly correlated labels. In this way, to find relevant features across multiple labels, the proposed method incorporates implicit and explicit label correlations and alleviates the negative influences of imperfect label information. Additionally, a manifold regularization is imposed to preserve the local feature structure in the latent space. Finally, to enhance the interpretability and to guide feature selection, we introduce sparse non-negative matrix regression with $L_{2,1}$-norm. We highlight the following contributions made by this paper:

- Using global and local label correlation for selecting discriminative features.
- Fusing label and feature information into a shared low-dimensional space for extracting label-feature relevance.
- Introducing a Nonnegative Matrix Factorization (NMF) (Lee & Seung, 1999) model that has inherent clustering and interpretability properties for selecting the most discriminative features.
- Adopting a graph regularization to guarantee the consistency between the original feature space and the latent space.
- Developing an effective optimization scheme to solve the MLFS-GLOCAL method.

The remainder of the paper is structured as shown below. Section 2 presents foundational concepts and related works. The details of the proposed model are presented in Section 3. Section 4 presents the experimental results. Finally, the conclusion and future works are provided in Section 5.

2. Preliminaries

In this paper, we use bold uppercase letters to signify matrices, such as matrix $\boldsymbol{A}$. When matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, $\boldsymbol{a}_i$ and $\boldsymbol{a}^{(j)}$ represent the ith column and the jth row of $\boldsymbol{A}$, respectively. Additionally, scalars are signified by lowercase letters like b, whereas vectors are represented by bold italicized lowercase letters like $\boldsymbol{b}$. $\boldsymbol{A}^{\top}$ and $\mathrm{Tr}(\boldsymbol{A})$ stand in for the transpose and trace of $\boldsymbol{A}$, respectively. $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} A_{i,j}^2}$ and $\|A\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{d} A_{i,j}^2}$ demonstrate the Frobenius norm and $L_{2,1}$-norm of matrix $\boldsymbol{A}$, respectively, where A_{ij} represents the (i,j)th entry in matrix $\boldsymbol{A}$. We define $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ as a d feature space with n instances, and $\boldsymbol{Y} \in \mathbb{R}^{n\times l}$ indicates the same instances in a l label space. If the ith sample has the j-label, then $Y_{i,j} = 1$; otherwise, $Y_{i,j} = 0$.

2.1. Related work

In recent years, there has been an increase in the prevalence of feature selection techniques for managing multi-label data, where each instance is associated with multiple labels. The research on multi-label feature selection has advanced quickly as more multi-label data have been studied. Through the research on the most advanced multi-label feature selection, the existing multi-label feature selection methods are mainly based on information-theoretic and embedded-based methods. Information-theoretic techniques use mutual information or conditional mutual information to extract the correlation between each candidate feature and each class label. As an illustration, the Max-Dependence and Min-Redundancy (MDMR) (Lin, Hu, Liu, & Duan, 2015) approach selects the best feature subset by increasing the feature dependence between features and labels while reducing feature redundancy. The following objective function shows how MDMR actually works:

$$J(\boldsymbol{f}_k) = \sum_{\boldsymbol{l}_i\in L} I(\boldsymbol{f}_k;\boldsymbol{l}_i) - \frac{1}{|\,S\,|}\sum_{\boldsymbol{f}_j\in S}\left\{ I(\boldsymbol{f}_k;\boldsymbol{f}_j) - \sum_{\boldsymbol{l}_i\in L} I(\boldsymbol{f}_k;\boldsymbol{l}_i|\boldsymbol{f}_j)\right\}, \tag{1}$$

where $\boldsymbol{f}_k, \boldsymbol{f}_j$, and S, stand for candidate feature, already-selected feature, and the already-selected feature subset, respectively. In contrast, $I(\boldsymbol{f}_k;\boldsymbol{l}_i)$ quantifies the feature dependence. The feature redundancy is measured by $I(\boldsymbol{f}_k;\boldsymbol{f}_j) - \sum_{\boldsymbol{l}_i\in L} I(\boldsymbol{f}_k;\boldsymbol{l}_i|\boldsymbol{f}_j)$. Similar to this, the SCLS feature selection approach proposed by Lee and Kim (2017) consists of the feature relevance term and scalable relevance evaluation.

$$J(\boldsymbol{f}_k) = \sum_{\boldsymbol{l}_i\in L} I(\boldsymbol{f}_k;\boldsymbol{l}_j) - \sum_{\boldsymbol{f}_j\in S}\frac{I(\boldsymbol{f}_k;\boldsymbol{f}_j)}{H(\boldsymbol{f}_k)}\sum_{\boldsymbol{l}_i\in L} I(\boldsymbol{f}_k;\boldsymbol{l}_j), \tag{2}$$

where the kth feature's entropy is denoted by $H(\boldsymbol{f}_k)$. Recently, the LRFS technique for multi-label feature selection was proposed (Zhang, Liu, and Gao, 2019). It is based on redundant labels. Labels are divided into independent and dependent categories by LRFS. Following is a presentation of the model:

$$\begin{aligned} J(\boldsymbol{f}_k) &= LR(\boldsymbol{f}_k;L) - \frac{1}{|S|}\sum_{\boldsymbol{f}_j\in S} I(\boldsymbol{f}_k;\boldsymbol{f}_j) \\ &= \sum_{\boldsymbol{l}_i\in L}\left\{\sum_{\boldsymbol{l}_i\neq \boldsymbol{l}_j,\boldsymbol{l}_j\in L} I(\boldsymbol{f}_k;\boldsymbol{l}_j|\boldsymbol{l}_i) - \frac{1}{|S|}\sum_{\boldsymbol{f}_j\in S} I(\boldsymbol{f}_k;\boldsymbol{f}_j)\right\}. \end{aligned} \tag{3}$$

Methods based on information theory disregard high-order interaction connections between features and labels. The significance of each individual feature or label is therefore a key factor in how effective these strategies are. Instead, multi-label feature selection embedded-based approaches emphasize the use of label correlations, using label correlations to choose the compact feature subset. In recent years, there have been several different sparse embedded-based feature selection techniques (Cai, Nie, & Huang, 2013; Jian et al., 2016; Nie, Huang, Cai, & Ding, 2010). As one of the basic methods, Nie et al. (2010) introduced the effective and Robust Feature Selection via the joint $L_{2,1}$-norm minimization (RFS) which can be formulated as follows:

$$\min_{\boldsymbol{W}} \sum_{i=1}^{n} \|\boldsymbol{x}_i\boldsymbol{W} - \boldsymbol{y}_i\|_2 + \gamma\|\boldsymbol{W}\|_{2,1} = \min_{\boldsymbol{W}} \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_{2,1} + \gamma\|\boldsymbol{W}\|_{2,1}, \tag{4}$$

where $\boldsymbol{W} \in \mathbb{R}^{d\times c}$ is a coefficient matrix, and the most discriminating features are chosen using $\|\boldsymbol{W}\|_{2,1}$.

Although RFS is a multi-class feature selection method, its framework is suitable for multi-label ones, and it is employed in numerous multi-label feature selection methods. For instance, Zhu et al. (2018) developed the RFS model for Missing Label Multi-Label Feature Selection (MLMLFS). This robust linear regression utilized a graph regularization based on the assumption that similar instances have similar labels. Additionally, a subset of discriminant features was selected by imposing an $L_{2,1}$, p-norm constraint with $0 < p \leq 1$. Since learning the multi-label regression models on a binary label matrix is challenging, recent researches attempt to utilize alternatives for label matrix. These studies can be categorized into two methods either using a latent label matrix or a pseudo-label matrix as the regression goal. The latent-based method known as MIFS stands for Multi-label Informed Feature Selection (Jian et al., 2016) and takes advantage of implicit label correlations to choose the most discriminating features. Additionally, MIFS takes into account the reduced low-dimensional label matrix to prevent exponential growth in the total number of features and labels. The objective function of MIFS is as follows:

$$\min_{\boldsymbol{W},\boldsymbol{V},\boldsymbol{B}} \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{V}\|_F^2 + \alpha\|\boldsymbol{Y} - \boldsymbol{V}\boldsymbol{B}\|_F^2 + \beta\mathrm{Tr}(\boldsymbol{V}^\top\boldsymbol{L}\boldsymbol{V}) + \gamma\|\boldsymbol{W}\|_{2,1}, \tag{5}$$

where $\boldsymbol{W}$ is a coefficient matrix of features and $\boldsymbol{X}$ is a feature matrix. The label matrix $\boldsymbol{Y}$'s coefficient matrix is denoted by the $\boldsymbol{V}$, while its latent semantics matrix is denoted by the $\boldsymbol{B}$. In other words, MIFS extracts common latent information from the feature and label matrices. This pioneer concept has been developed in some state-of-the-art multi-label feature selection methods. The MIFS method decomposes the multi-label matrix into two-factor matrices, which contain entries of mixed signs. As a result, interpreting them can be challenging. Braytee, Liu, Catchpoole, and Kennedy (2017) presented CMFS, a multi-label feature selection approach that maps feature and label matrices into a shared low-dimensional space by a joint tri-factorization and local structure preservation. Also, this model attempts to maximize the dependence between latent correlations that are extracted from feature and label matrices.

$$\begin{aligned}\min_{\boldsymbol{V},\boldsymbol{L},\boldsymbol{Q},\boldsymbol{P},\boldsymbol{B}} &\|\boldsymbol{X} - \boldsymbol{V}\boldsymbol{L}\boldsymbol{Q}\|_F^2 + \alpha\|\boldsymbol{Y} - \boldsymbol{V}\boldsymbol{P}\boldsymbol{B}\|_F^2 + \beta\|\boldsymbol{L} - \boldsymbol{P}\|_F^2 \\ &+ \epsilon\mathrm{Tr}(\boldsymbol{R}(\boldsymbol{V}\boldsymbol{P}\boldsymbol{B})^\top\boldsymbol{V}\boldsymbol{P}\boldsymbol{B}) + \gamma\|\boldsymbol{Q}\|_{2,1} \\ &\text{s.t.}\quad \boldsymbol{V},\boldsymbol{L},\boldsymbol{Q},\boldsymbol{P},\boldsymbol{B} \geq 0.\end{aligned} \tag{6}$$

Shared Common Mode Feature Selection (SCMFS) (Hu, Li, Gao, Zhang, & Hu, 2020) is another extension of MIFS that similarly recovers the shared latent information between feature space and label space in an NMF-based model. To verify the matrix regression term, SCMFS adds a decoder factorization term on the feature matrix.

$$\begin{aligned}&\min_{\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}} \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{V}\|_F^2 + \alpha\|\boldsymbol{X} - \boldsymbol{V}\boldsymbol{Q}\|_F^2 + \beta\|\boldsymbol{Y} - \boldsymbol{V}\boldsymbol{B}\|_F^2 + \gamma\|\boldsymbol{W}\|_{2,1}, \\ &\text{s.t. } \boldsymbol{W},\boldsymbol{V},\boldsymbol{Q},\boldsymbol{B} \geq 0.\end{aligned} \tag{7}$$

More recently, Gao, Li, and Hu (2023) proposed the Shared Structure Feature Selection (SSFS) that changed the regression term (encoder structure) in the MIFS to a factorization term (decoder structure), and similar to the SCMFS, optimized its cost function in an NMF-based model.

$$\begin{aligned}&\min_{\boldsymbol{V},\boldsymbol{Q},\boldsymbol{M}} \|\boldsymbol{X} - \boldsymbol{V}\boldsymbol{Q}^\top\|_F^2 + \alpha\|\boldsymbol{Y} - \boldsymbol{V}\boldsymbol{M}\|_F^2 + \beta\mathrm{Tr}(\boldsymbol{V}^\top\boldsymbol{L}\boldsymbol{V}) + \gamma\|\boldsymbol{Q}\|_{2,1}, \\ &\text{s.t.}\quad \boldsymbol{V},\boldsymbol{M},\boldsymbol{Q} \geq 0.\end{aligned} \tag{8}$$

There are various pseudo-label-based methods in the MLFS literature that make a connection between the label matrix and its alternative in different ways. For example, Huang and Wu (2021) proposed an MLFS method with manifold regularization and dependence maximization (MRDM). This method replaces the label space with a manifold embedding. This embedding is constrained through manifold regularization. In addition, they use Hilbert Schmidt Independence Criterion (HSIC) as a regularization to maximize the dependence between the manifold embedding and the label matrix.

$$\min_{\boldsymbol{W},\boldsymbol{Z}^\top\boldsymbol{Z}=\boldsymbol{I}} \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Z}\|_F^2 + \alpha\mathrm{Tr}(\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{Z}) - \beta\mathrm{Tr}(\boldsymbol{H}\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{H}\boldsymbol{Y}\boldsymbol{Y}^\top) + \gamma\|\boldsymbol{W}\|_{2,1}. \tag{9}$$

Fan, Liu, Liu et al. (2021) proposed a dual manifold regularized framework that embeds the feature matrix into two latent spaces (label space and cluster space). Also, this model utilizes a regularization to maximize the dependence between the manifold embedding and the label matrix. The implicit label correlation is exploited by preserving the global and local structural information.

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{V},\boldsymbol{F},\boldsymbol{Q}} &\|\boldsymbol{X}^\top\boldsymbol{W} - \boldsymbol{V}\|_{2,1} + \alpha\|\boldsymbol{W}\|_{2,1} + \beta\|\boldsymbol{W} - \boldsymbol{F}\boldsymbol{Q}\|_F^2 + \gamma\mathrm{Tr}(\boldsymbol{F}^\top\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top\boldsymbol{F}) \\ &+ \lambda(\mathrm{Tr}[(\boldsymbol{V} - \boldsymbol{Y})^\top\boldsymbol{E}(\boldsymbol{V} - \boldsymbol{Y})] + \mathrm{Tr}(\boldsymbol{V}^\top\boldsymbol{L}\boldsymbol{V})), \\ &\text{s.t.}\quad \boldsymbol{V} \geq 0, \boldsymbol{W}^\top\boldsymbol{W} = \boldsymbol{I}, \boldsymbol{F}^\top\boldsymbol{F} = \boldsymbol{I}.\end{aligned} \tag{10}$$

Similarly, Fan, Liu, Weng et al. (2021) proposed a robust method that integrates multi-label feature selection and the local discriminant model. It clusters the feature weight matrix by incorporating discriminative information to exploit implicit label correlation.

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{b},\boldsymbol{L},\boldsymbol{P}} &\|\boldsymbol{X}\boldsymbol{W} + \boldsymbol{1}_n\boldsymbol{b}^\top - \boldsymbol{Y}\|_{2,1} + \alpha\|\boldsymbol{W} - \boldsymbol{L}\boldsymbol{P}\|_F^2 \\ &+ \beta\mathrm{Tr}(\boldsymbol{L}^\top\boldsymbol{M}\boldsymbol{L}) + \gamma\|\boldsymbol{W}\|_{2,1}, \quad \text{s.t.}\quad \boldsymbol{L}^\top\boldsymbol{L} = \boldsymbol{I}.\end{aligned} \tag{11}$$

Zhang and Ma (2022) presented an NMDG or nonnegative multi-label feature selection with dynamic graph constraints. In NMDG, the pseudo-label matrix is trained using label manifolds and linear regression. Additionally, the feature manifold is merged with the pseudo-label to create the dynamic graph Laplacian matrix, which is then utilized to constrain the learning of the feature weight matrix.

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{b},\boldsymbol{F}} &\|\boldsymbol{X}\boldsymbol{W} + \boldsymbol{1}_n\boldsymbol{b}^\top - \boldsymbol{F}\|_F^2 + \alpha\mathrm{Tr}(\boldsymbol{F}^\top\boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{F}) + \beta\mathrm{Tr}(\boldsymbol{W}\boldsymbol{L}_{\boldsymbol{F}^\top}\boldsymbol{W}^\top) \\ &+ \gamma\mathrm{Tr}(\boldsymbol{W}^\top\boldsymbol{L}_{\boldsymbol{X}^\top}\boldsymbol{W}), \quad \text{s.t.}\quad (\boldsymbol{W},\boldsymbol{F}) \geq 0.\end{aligned} \tag{12}$$

Besides the multi-label feature selection methods mentioned, there are some specific multi-label learning methods that explore high-order label correlations. As a pioneering work, Zhu, Kwok, and Zhou (2017) presented the GLOCAL multi-label correlation learning strategy, which at the same time retrieves missing labels, trains the classifier, and utilizes global and local label correlations by optimizing the label manifolds and learning a latent label representation. Zhao, Gao, Lu, and Sun (2022) introduced a multi-label learning method called LSGL. Considering the suppositions of global label consistency and local label smoothness, this method learns a label correlation matrix. LSGL attempts to extract label correlation from the global and local perspectives in a self-representation model and a local data structure framework, respectively. Kumar and Rastogi (2022) presented the Transformation of Low-Rank Label Subspace for Multi-label Learning with Missing Labels (LRMML). The framework extracts global label correlation by a self-representation model and models the transformation of the low-rank label subspace, which is used to restore missing labels and train the classifier.

3. Proposed method

This section introduces the Multi-Label Feature Selection with Global and Local Label Correlation (MLFS-GLOCAL), which uses both global and local label correlations for selecting the relevant and non-redundant features. The method's success is attributed to four primary factors: (1) To provide a more insightful feature-label representation, it extracts the low-rank structure from the label and feature matrices, which also provides an implicit label correlation (Section 3.1); (2) It imposes a penalty that preserves the smoothness of local mapping in the shared latent space. (Section 3.2); (3) It can leverage information from all labels by taking into account both global and local label correlations.

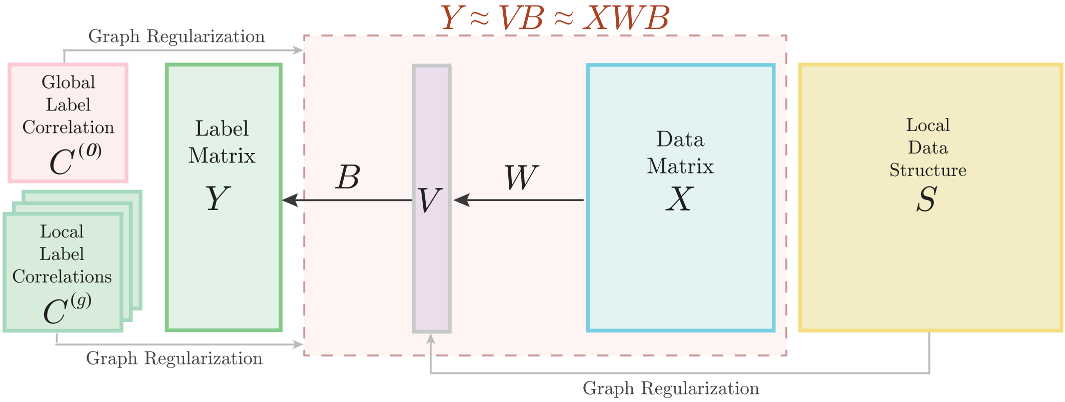


Fig. 1. Schematic representation of the proposed multi-label feature selection model (MLFS-GLOCAL).

(Section 3.3); (Section 3.4) It combines the aforementioned into a single joint learning problem and employs an effective alternating minimization approach for optimization. Fig. 1 provides a schematic representation of the MLFS-GLOCAL model.

3.1. Shared latent space

We define the feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, and demonstrate the ground-truth label matrix $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l\} \subseteq \{0,1\} \in \mathbb{R}^{n\times l}$, where $Y_{ij} = 1$ if ith instance would have the jth label, and otherwise $Y_{ij} = 0$. In multi-label problems, the labels are associated, hence it is common to assume that the label matrix is low-rank. The label matrix $\boldsymbol{Y}$ is sparse, binary-valued, and high dimensional, therefore learning from this matrix is difficult. Let the rank of $\boldsymbol{Y}$ be $k < \min(n, l)$. $\boldsymbol{Y}$ can be factorized into two smaller matrices as follows:

$$\boldsymbol{Y} \simeq \boldsymbol{V}\boldsymbol{B}, \tag{13}$$

where $\boldsymbol{V} \in \mathbb{R}^{n\times k}$ and $\boldsymbol{B} \in \mathbb{R}^{k\times l}$. Intuitively, $\boldsymbol{V}$ reflects the latent labels, which are more compact and semantically abstract than the original labels. while $\boldsymbol{B}$ reflects how the original labels are correlated to the latent labels. The significant information of the original label matrix $\boldsymbol{Y}$ is coded in the low-dimensional matrix $\boldsymbol{V}$, which also reduces the label matrix's unfavorable data. Matrices $\boldsymbol{V}$ and $\boldsymbol{B}$ can be obtained by minimizing the label representation error $\|\boldsymbol{Y} - \boldsymbol{V}\boldsymbol{B}\|_F^2$. The latent label representation, denoted as $\boldsymbol{V}$, is a low-dimensional, real-valued matrix that contains dense information. It is comparatively easier to learn a continuous mapping from the feature space to the latent label space than to the original label space (Zhu et al., 2017). Similar features plan to have similar labels, which is one important assumption regarding correlations between features and labels. As a result, the shared information between the feature space and the label space should be consistent. We consider $\boldsymbol{V}$ to be a shared factor matrix between the feature matrix and the label matrix. We learn a matrix $\boldsymbol{W} \in \mathbb{R}^{d\times k}$ to map instances to the latent space. By decreasing the feature representation error $\|\boldsymbol{V} - \boldsymbol{X}\boldsymbol{W}\|_F^2$, the feature weight matrix $\boldsymbol{W}$ would be learnt. By integrating label representation and feature representation losses in a Nonnegative Matrix Factorization framework (Lee & Seung, 1999), we learn shared latent space through the following objective function:

$$\min_{\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{V}\boldsymbol{B}\|_F^2 + \|\boldsymbol{V} - \boldsymbol{X}\boldsymbol{W}\|_F^2, \quad \text{s.t.} \quad \boldsymbol{V},\boldsymbol{B},\boldsymbol{W} \geq 0. \tag{14}$$

3.2. Local structure preservation

In many machine learning applications, the feature space can be high-dimensional and noisy, making it challenging to extract meaningful information from the data. One popular approach is to transform the original feature space into a lower-dimensional latent space, where the underlying structure of the data can be more easily captured. However, simply projecting the data into a lower-dimensional space can result in a loss of information and may not accurately capture the complex relationships among the features. In order to solve this issue, we use the graph regularization technique (Cai, He, Han, & Huang, 2010) to provide a suitable transformation matrix $\boldsymbol{V}$ that ensures the coherence between the initial feature space and the latent structure space. According to the underlying principle of this regularization, the closer correlation between two instances in the feature matrix $\boldsymbol{X}$ denotes the closer correlation between the two corresponding latent feature variables $\boldsymbol{v}^{(i)}$ and $\boldsymbol{v}^{(j)}$ in the latent structure. Specifically, we utilize a general graph regularization term to encourage the locality preservation in the latent space, ensuring that the latent variables accurately reflect the structure of the original features. The graph regularization term can be expressed as:

$$\min_{\boldsymbol{V}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)}\|^2 S_{i,j} = \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{D}\boldsymbol{V}) - \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{S}\boldsymbol{V}) = \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{L}\boldsymbol{V}), \tag{15}$$

where $\boldsymbol{D}$ is a matrix diagonal, $\boldsymbol{S}$ indicates a symmetric affinity matrix, and $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$ is the graph Laplacian matrix. Integrating the above terms into our model, we achieve the following function:

$$\min_{\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}} \|\boldsymbol{Y}-\boldsymbol{V}\boldsymbol{B}\|_F^2+\|\boldsymbol{V}-\boldsymbol{X}\boldsymbol{W}\|_F^2+\lambda_1 \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{L}\boldsymbol{V}), \quad \text{s.t.} \quad \boldsymbol{V},\boldsymbol{B},\boldsymbol{W} \geq 0, \tag{16}$$

where λ_1 is the local structure parameter.

3.3. Global and local label correlation

To effectively utilize the information of multiple labels, label correlations must be incorporated. In the proposed method, we regularize the proposed model using high-order label correlation. In this direction, the coexistence of global and local label correlations should be noted. To consider both of them, we introduce label manifold regularizers in this part. The concept behind the global manifold regularizer is derived from the instance-level manifold regularizer, as outlined in (15). In particular, if two labels are highly correlated, their corresponding classifier outputs should be more similar, and conversely, less correlated labels should produce less similar classifier outputs. In other words, label correlations lead to similar classifier outputs. This paper employs cosine similarity to quantify the global label correlation, denoted as $\boldsymbol{C} \in \mathbb{R}^{l\times l}$, which is calculated as follows:

$$C_{ij} = \frac{\boldsymbol{y}_i^\top \boldsymbol{y}_j}{\|\boldsymbol{y}_i\|\|\boldsymbol{y}_j\|}, \tag{17}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ indicate the ith and jth label vectors for all instances.

In the basic model (16), the label predicted for sample $\boldsymbol{x}$ is $\boldsymbol{f}(\boldsymbol{x})$, where $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{W}\boldsymbol{B}$. Let $\boldsymbol{f} = \{f_1, \ldots, f_l\}$, where $f_j(\boldsymbol{x})$ is the jth anticipated label for sample $\boldsymbol{x}$. As a result, predictions for all n instances are recorded in the prediction matrix $\boldsymbol{F} \in \mathbb{R}^{n\times l}$, where $\boldsymbol{F} = \boldsymbol{XWB}$ contains predictions. If the ith and jth labels are more correlated, their corresponding predict labels $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ should be more similar to each other. The label manifold regularization is defined similarly to the instance-level manifold regularization (15) as follows:

$$\min_{\boldsymbol{F}} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \|\boldsymbol{f}_i - \boldsymbol{f}_j\|^2 C_{i,j} = \mathrm{Tr}(\boldsymbol{FAF}^\top) - \mathrm{Tr}(\boldsymbol{FCF}^\top) = \mathrm{Tr}(\boldsymbol{FPF}^\top), \quad (18)$$

where $\boldsymbol{A}$ is a diagonal matrix as $A_{ii} = \sum_{j=1}^{l} C_{ij}$ and $\boldsymbol{P} = \boldsymbol{A} - \boldsymbol{C}$. By minimizing (18), $\|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2$ will be small. The manifold regularizer in (18) written as $\mathrm{Tr}(\boldsymbol{FPF}^\top)$, where $\boldsymbol{P} = \boldsymbol{A} - \boldsymbol{C}$ is the label Laplacian matrix and $\boldsymbol{F} = \boldsymbol{XWB}$ is the classifier output matrix. Since label correlations can differ across different local regions, we propose the local manifold regularizer. We assume that the dataset $\boldsymbol{X}$ is divided into g groups $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_g\}$, where $\boldsymbol{X}_m \in \mathbb{R}^{n_m \times d}$ matrix has $\boldsymbol{n}_m$ instances. By using clustering or domain knowledge, such as networks and gene pathways (Chuang, Lee, Liu, Lee, & Ideker, 2007; Hajiveiseh, Seyedi, & Akhlaghian Tab, 2024; Subramanian et al., 2005) in bioinformatics applications, it is possible to acquire this partitioning. Assume $\boldsymbol{C}_m \in \mathbb{R}^{l\times l}$ is the local label correlation matrix of a group m and that $\boldsymbol{Y}_m$ is the label submatrix in $\boldsymbol{Y}$ corresponding to $\boldsymbol{X}_m$. The same as (17), we calculate the local label correlations as follows:

$$C_{i,j}^{(m)} = \frac{{\boldsymbol{y}_i^{(m)}}^\top \boldsymbol{y}_j^{(m)}}{\|\boldsymbol{y}_i^{(m)}\|\|\boldsymbol{y}_j^{(m)}\|}, \; m \in \{1, \ldots, g\}. \quad (19)$$

Analogous to the global label correlations, we motivate the classifier outputs to be similar on the correlated labels as follows:

$$\min_{\boldsymbol{F}} \sum_{m=1}^{g} \frac{n_m}{n} \sum_{i=1}^{l}\sum_{j=1}^{l} \|\boldsymbol{f}_i^{(m)} - \boldsymbol{f}_j^{(m)}\|^2 C_{i,j}^{(m)} \quad (20)$$
$$= \sum_{m=1}^{g} \frac{n_m}{n}[\mathrm{Tr}(\boldsymbol{F}_m \boldsymbol{A}_m \boldsymbol{F}_m^\top) - \mathrm{Tr}(\boldsymbol{F}_m \boldsymbol{C}_m \boldsymbol{F}_m^\top)] = \sum_{m=1}^{g} \frac{n_m}{n}\mathrm{Tr}(\boldsymbol{F}_m \boldsymbol{P}_m \boldsymbol{F}_m^\top),$$

where $\boldsymbol{F}_m = \boldsymbol{X}_m\boldsymbol{WB}$ is the classifier output matrix for group m and $\boldsymbol{P}_m$ is the Laplacian matrix of $\boldsymbol{C}_m$. To cover the cluster imbalance, we scale each local label correlation regularization by a coefficient n_m/n. Problem (16) now has the following optimization problem after the addition of global and local manifold regularizers (18) and (20):

$$\min_{\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{VB}\|_F^2 + \|\boldsymbol{V} - \boldsymbol{XW}\|_F^2 + \lambda_1 \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{LV}) \quad (21)$$
$$+ \lambda_2[\mathrm{Tr}(\boldsymbol{FPF}^\top) + \sum_{m=1}^{g}\frac{n_m}{n}\mathrm{Tr}(\boldsymbol{F}_m\boldsymbol{P}_m\boldsymbol{F}_m^\top)], \quad \text{s.t.} \quad \boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W} \geq 0,$$

where λ_2 indicates the impact of global and local label correlation on the objective function. Finally, our function uses the $L_{2,1}$-norm, which has been shown to be beneficial for feature selection. As a result, the objective function is set up as follows:

$$\min_{\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{VB}\|_F^2 + \|\boldsymbol{V} - \boldsymbol{XW}\|_F^2 + \lambda_1 \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{LV}) + \lambda_2[\mathrm{Tr}(\boldsymbol{FPF}^\top) \quad (22)$$
$$+ \sum_{m=1}^{g}\frac{n_m}{n}\mathrm{Tr}(\boldsymbol{F}_m\boldsymbol{P}_m\boldsymbol{F}_m^\top)] + \lambda_3\|\boldsymbol{W}\|_{2,1}, \quad \text{s.t.} \quad \boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W} \geq 0,$$

where the sparsity of $\boldsymbol{W}$ on rows is ensured by using the $L_{2,1}$-norm. Adjusting the objective function's sparsity is done using the λ_3 parameter.

3.4. Optimization

Function (22) contains the $L_{2,1}$-norm regularization term. Since it is not smooth, a direct solution is not possible. Additionally, when variables $\boldsymbol{V}$, $\boldsymbol{B}$, and $\boldsymbol{W}$ are considered together, it is non-convex. In other words, the Hessian matrix created from the partial derivatives of the objective function at the second degree is not a matrix that is positively semi-definite. The objective function must be optimized such that it is convex for each iteration by fixing any three of the variables and updating the remaining one. The phrase $\|\boldsymbol{W}\|_{2,1}$ is further relaxed using $\mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{D}_w \boldsymbol{W})$, $\boldsymbol{D}_w$ is a diagonal matrix in this case (Hu et al., 2020). The algorithm's iterative updating feature is available. The $\boldsymbol{D}_w$ element is $D_{ii} = 1/(\|\boldsymbol{w}_i\| + \epsilon), (\epsilon \leftarrow 0)$, where ϵ stops the non-differentiable problem's disturbance. As a result, the objective function (22) can be rewritten as follows:

$$\min_{\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{VB}\|_F^2 + \|\boldsymbol{V} - \boldsymbol{XW}\|_F^2 + \lambda_1 \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{LV}) + \lambda_2[\mathrm{Tr}(\boldsymbol{FPF}^\top) \quad (23)$$
$$+ \sum_{m=1}^{g}\frac{n_m}{n}\mathrm{Tr}(\boldsymbol{F}_m\boldsymbol{P}_m\boldsymbol{F}_m^\top)] + \lambda_3(\boldsymbol{W}^\top \boldsymbol{D}_w \boldsymbol{W}), \quad \text{s.t.} \quad \boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W} \geq 0.$$

We introduce Lagrangian multipliers to incorporate nonnegative constraint conditions into the function. $\boldsymbol{\Phi}, \boldsymbol{\Psi}$, and $\boldsymbol{\Omega}$ to restrict $\boldsymbol{V}$, $\boldsymbol{B}$, and $\boldsymbol{W}$ respectively, where $\boldsymbol{\Phi} \in \mathbb{R}^{n\times k}$, $\boldsymbol{\Psi} \in \mathbb{R}^{k\times l}$, $\boldsymbol{\Omega} \in \mathbb{R}^{d\times k}$, As a result, the function (23) is equivalent to the following function:

$$\min_{\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{VB}\|_F^2 + \|\boldsymbol{V} - \boldsymbol{XW}\|_F^2 + \lambda_1 \mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{LV}) + \lambda_2[\mathrm{Tr}(\boldsymbol{FPF}^\top) \quad (24)$$
$$+ \sum_{m=1}^{g}\frac{n_m}{n}\mathrm{Tr}(\boldsymbol{F}_m\boldsymbol{P}_m\boldsymbol{F}_m^\top)] + \lambda_3\mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{D}_w \boldsymbol{W})$$
$$- \mathrm{Tr}(\boldsymbol{\Phi V}^\top) - \mathrm{Tr}(\boldsymbol{\Psi B}^\top) - \mathrm{Tr}(\boldsymbol{\Omega W}^\top).$$

Given an arbitrary matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_F^2 = \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{A})$. Therefore, the function (24) becomes:

$$C = \mathrm{Tr}[(\boldsymbol{Y} - \boldsymbol{VB})^\top(\boldsymbol{Y} - \boldsymbol{VB})] + \mathrm{Tr}[(\boldsymbol{V} - \boldsymbol{XW})^\top(\boldsymbol{V} - \boldsymbol{XW})] \quad (25)$$
$$+ \lambda_1\mathrm{Tr}(\boldsymbol{V}^\top \boldsymbol{LV}) + \lambda_2[\mathrm{Tr}(\boldsymbol{FPF}^\top) + \sum_{m=1}^{g}\frac{n_m}{n}\mathrm{Tr}(\boldsymbol{F}_m\boldsymbol{P}_m\boldsymbol{F}_m^\top)]$$
$$+ 2\lambda_3\mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{D}_w \boldsymbol{W}) - \mathrm{Tr}(\boldsymbol{\Phi V}^\top) - \mathrm{Tr}(\boldsymbol{\Psi B}^\top) - \mathrm{Tr}(\boldsymbol{\Omega W}^\top),$$

where $\boldsymbol{F} = \boldsymbol{XWB}$ and $\boldsymbol{F}_m = \boldsymbol{X}_m\boldsymbol{WB}$, $\forall m \in \{1, 2, \ldots, g\}$.

The partial derivatives of function (25) with respect to the variables $\boldsymbol{V}$, $\boldsymbol{B}$, and $\boldsymbol{W}$ are:

$$\frac{\partial C}{\partial \boldsymbol{W}} = -\boldsymbol{X}^\top\boldsymbol{V} + \boldsymbol{X}^\top\boldsymbol{XW} + \lambda_2[\boldsymbol{X}^\top\boldsymbol{XWBPB}^\top \quad (26)$$
$$+ \sum_{m=1}^{g}\frac{n_m}{n}\boldsymbol{X}_m^\top\boldsymbol{X}_m\boldsymbol{WBP}_m\boldsymbol{B}^\top] + 2\lambda_3\boldsymbol{D}_w\boldsymbol{W} - \boldsymbol{\Omega},$$

$$\frac{\partial C}{\partial \boldsymbol{B}} = -\boldsymbol{V}^\top\boldsymbol{Y} + \boldsymbol{V}^\top\boldsymbol{VB} + \lambda_2[\boldsymbol{W}^\top\boldsymbol{X}^\top\boldsymbol{XWBP}$$
$$+ \sum_{m=1}^{n}\frac{n_m}{n}\boldsymbol{W}^\top\boldsymbol{X}_m^\top\boldsymbol{X}_m\boldsymbol{WBP}_m] - \boldsymbol{\Phi}, \quad (27)$$

and

$$\frac{\partial C}{\partial \boldsymbol{V}} = \boldsymbol{V} - 2\boldsymbol{XW} - 2\boldsymbol{YB}^\top + \boldsymbol{VBB}^\top + \lambda_1\boldsymbol{LV} - \boldsymbol{\Psi}. \quad (28)$$

By setting the partial derivatives (26), (27), and (28) to zero, we can find critical points of the function. In accordance with the Karush–Kuhn–Tucker condition, we set $\boldsymbol{W} \odot \boldsymbol{\Omega} = \boldsymbol{0}$, $\boldsymbol{B} \odot \boldsymbol{\Phi} = \boldsymbol{0}$, and $\boldsymbol{V} \odot \boldsymbol{\Psi} = \boldsymbol{0}$, that are fixed point equations that the solution must satisfy at convergence. By solving these equations and with respect to $\boldsymbol{P} = \boldsymbol{A} - \boldsymbol{C}$ and $\boldsymbol{P}_{\boldsymbol{m}} = \boldsymbol{A}_m - \boldsymbol{C}_m$, we derive the following updating rules for the $\boldsymbol{W}$, $\boldsymbol{B}$, and $\boldsymbol{V}$:

$$\boldsymbol{W} \leftarrow \boldsymbol{W}$$
$$\odot \frac{\boldsymbol{X}^\top\boldsymbol{V} + \lambda_2[(\boldsymbol{X}^\top\boldsymbol{XWBCB}^\top) + \sum_{m=1}^{g}\frac{n_m}{n}(\boldsymbol{X}_m^\top\boldsymbol{X}_m\boldsymbol{WBC}_m\boldsymbol{B}^\top)]}{\boldsymbol{X}^\top\boldsymbol{XW} + \lambda_2(\boldsymbol{X}^\top\boldsymbol{XWBAB}^\top) + \sum_{m=1}^{g}\frac{n_m}{n}(\boldsymbol{X}_m^\top\boldsymbol{X}_m\boldsymbol{WBA}_m\boldsymbol{B}^\top) + \lambda_3(\boldsymbol{D}_w\boldsymbol{W})}, \quad (29)$$

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \odot \frac{\boldsymbol{V}^\top\boldsymbol{Y} + \lambda_2[(\boldsymbol{W}^\top\boldsymbol{X}^\top\boldsymbol{XWBC}) + \sum_{m=1}^{g}\frac{n_m}{n}(\boldsymbol{W}^\top\boldsymbol{X}_m^\top\boldsymbol{X}_m\boldsymbol{WBC}_m)]}{\boldsymbol{V}^\top\boldsymbol{VB} + \lambda_2[(\boldsymbol{W}^\top\boldsymbol{X}^\top\boldsymbol{XWBA}) + \sum_{m=1}^{g}\frac{n_m}{n}(\boldsymbol{W}^\top\boldsymbol{X}_m^\top\boldsymbol{X}_m\boldsymbol{WBA}_m)]}, \quad (30)$$

Algorithm 1 Multi-label Feature Selection with Global and Local label correlation
(MLFS-GLOCAL)

```
Input: Feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and Label matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times l}$,
Regularization parameters $\lambda_1, \lambda_2$, and $\lambda_3$, and latent factor $k$;
Output: Feature score $s_i = \|\boldsymbol{w}_i\|, \forall i \in \{1, 2, \ldots, d\}$.

Calculate the similarity matrix $\boldsymbol{S}$;
Calculate the Global label correlation $\boldsymbol{C}$ and Local label correlation $\boldsymbol{C}_m$;
Initialize $\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{B}$ randomly; $t = 0$;
while $t <$ MaxIteration do
    Update $D_{ii} \leftarrow \frac{1}{\|\boldsymbol{w}_i\|+\epsilon}$;
    Update $\boldsymbol{W}$ by (29);
    Update $\boldsymbol{B}$ by (30);
    Update $\boldsymbol{V}$ by (31);
    $t = t + 1$;
end while
Return $\boldsymbol{W}$;
Evaluate the feature score by $s_i = \|\boldsymbol{w}_i\|$.
```

Algorithm 2 Evaluation of Multi-label Feature Selection with ML–kNN Classifier

```
Input: train feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, test feature matrix $\boldsymbol{X}' \in \mathbb{R}^{n'\times d}$,
train label matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times l}$, test label matrix $\boldsymbol{Y}' \in \mathbb{R}^{n'\times l}$, feature scores
$\boldsymbol{s} = \{\|\boldsymbol{w}_i\|\}$, ML–kNN parameter $k_{nn}$;
Output: Evaluation results for multi-label feature selection.

▷ Filter Features:
Sort feature scores in descending order: $\boldsymbol{s} \leftarrow$ SortDescending($\boldsymbol{s}$);
Select the top $p$ features based on a specific number: $\boldsymbol{X}_s \leftarrow$
    SelectTopFeatures($\boldsymbol{X}, \boldsymbol{s}, p$) and $\boldsymbol{X}'_s \leftarrow$ SelectTopFeatures($\boldsymbol{X}', \boldsymbol{s}, p$);
▷ ML-kNN Classifier:
Train ML–kNN classifier using filtered train set $\boldsymbol{X}_s$ and corresponding labels $\boldsymbol{Y}$:
    model ← ML–kNN($\boldsymbol{X}_s$,$\boldsymbol{Y}$, $k_{nn}$);
Predict the prediction labels $\boldsymbol{P}'$ and scores $\boldsymbol{F}'$ on the test set $\boldsymbol{X}'_s$
    using the trained ML–kNN classifier: $[\boldsymbol{P}', \boldsymbol{F}'] \leftarrow$ model($\boldsymbol{X}'_s$);
▷ Evaluation Metrics:
Evaluate prediction labels $\boldsymbol{P}'$ and scores $\boldsymbol{F}'$ using test label matrix
    $\boldsymbol{Y}' \in \mathbb{R}^{n'\times l}$ and various evaluation metrics.
Return Evaluation results for multi-label feature selection.
```

and

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}\boldsymbol{W} + \boldsymbol{Y}\boldsymbol{B}^\top + \lambda_1 \boldsymbol{S}\boldsymbol{V}}{\boldsymbol{V} + \boldsymbol{V}\boldsymbol{B}\boldsymbol{B}^\top + \lambda_1 \boldsymbol{D}\boldsymbol{V}}. \tag{31}$$

MLFS-GLOCAL, which ranks all the features using $\|\boldsymbol{w}_i\|_2, (i = 1, \ldots, d)$ in descending order, allowing us to obtain the top-k features. Algorithm 1 outlines the detailed solution for the MLFS-GLOCAL model, while Algorithm 2 evaluates the proposed feature selection Algorithm 1 using the achieved feature scores $\boldsymbol{s} = \{\|\boldsymbol{w}_i\|_2\}$ and the ML-kNN classifier.

3.5. Complexity analysis

In the MLFS-GLOCAL Algorithm 1, the computational complexity is divided into two phases. During the pre-optimization stage, the construction of the similarity matrix $\boldsymbol{S}$, global correlation matrix $\boldsymbol{C}$, and local correlation matrices $\boldsymbol{C}_m$ are accomplished with the time complexities of $O(dn \log n)$, $O(nl \log l)$, and $O(n_m l \log l)$, respectively. In the iterative stage, characterized by t iterations, the update of the feature weight matrix $\boldsymbol{D}_w$ takes place in $O(dkt)$ time. Moreover, the update of $\boldsymbol{W}$, incorporating the constant calculation of $\boldsymbol{X}^\top\boldsymbol{X}$ once, has a complexity of $O(d^2kt + dnkt + d^2n)$, while the update of $\boldsymbol{V}$ follows a complexity of $O(dnkt + n^2kt)$, and the complexity of $\boldsymbol{B}$ is $O(nklt + nk^2t + d^2kt)$. Considering the number of features d, the overall computational complexity of Algorithm 1 depends on the interaction between d and the sample size n. In scenarios where d is significantly smaller or comparable to n, the dominating term is n^2kt, resulting in an overall complexity of $O(n^2kt)$. Conversely, for larger d compared to n, the dominating term becomes d^2n, resulting in an overall complexity of $O(d^2n)$.

Table 1
The detailed information of the real-world datasets.

Dataset	#Instance	#Feature	#Label
Arts	5000	462	26
Birds	645	260	19
Business	5000	438	30
Computers	5000	681	33
Corel5k	5000	499	374
Education	5000	550	33
Entertainment	5000	640	21
Health	5000	612	32
Recreation	5000	606	22
Reference	5000	793	33
Scene	2407	294	6
Science	5000	743	40
Social	5000	1047	39
Society	5000	636	27

4. Experimental study

This section conducts a comprehensive evaluation to assess the MLFS-GLOCAL model on 14 real-world multi-label benchmark datasets, using six diverse evaluation metrics. The proposed model is compared to nine well-known and state-of-the-art feature selection methods. All of our tests are run on an Intel Core (TM) i7-9700K with a 3.6 GHz processor and 32 GB RAM.

4.1. Datasets

In our studies, we employed 14 datasets from Mulan Library's multi-label text and image classification. The multi-label Yahoo datasets refer to a collection of datasets that are used for multi-label classification tasks. These datasets were originally released by Yahoo Labs and consist of a large number of text documents that have been annotated with multiple labels. Each datasets includes a training set and a test set that each comprises 2000 and 3000 documents, respectively (Doquire & Verleysen, 2013). The Corel5k multi-label dataset is a collection of images used for multi-label classification tasks. It consists of 5000 images from 374 different categories, with each image having multiple labels assigned to it. Additionally, the Scene dataset, another image dataset, comprises 2407 images categorized into 6 different groups. The Birds dataset, consisting of 645 audio instances and 19 labels. These datasets have been widely used in research for developing and evaluating multi-label classification algorithms. Table 1 describes the specifics of each benchmark dataset.

4.2. Evaluation metrics

To examine the performance of all competition methods, the Multi-Label kNN (ML-kNN) algorithm (Zhang & Zhou, 2007) is defined as a benchmark classifier. The ML-kNN is frequently used for classification in multi-label feature selection approaches (Hu et al., 2020; Jian et al., 2016; Liu, Lin, Wu and Wang, 2018; Zhang, Luo, Li, Zhou, & Li, 2019) because of its interpretability and simplicity. We set $p = 10$ for the number of nearest neighbors. Furthermore, we use six commonly used assessment criteria, including Micro-F1, Macro-F1, Average Precision, Ranking Loss, Hamming Loss, and Coverage Error. These assessment metrics' definitions are as follows:

- **Macro-F1** and **Micro-F1** both are on the measure's foundation of F1-measure. The evaluation metric directly uses F-measure averaging to rate the precision of the predictions made by the classifier label set.

$$Micro-F1=\frac{\sum_{i=1}^{l}2TP_i}{\sum_{i=1}^{l}(2TP_i+FP_i+FN_i)}, \tag{32}$$

and

$$Macro-F1=\sum_{i=1}^{l}\frac{2TP_i}{2TP_i+FP_i+FN_i}, \tag{33}$$

where T and F are True and False, respectively; P and N are Positive and Negative, respectively; so TP, TN, FP, and FN are the number of combinations of T, F, P, and N, respectively.

- **Average precision** determines the percentage of labels that are more relevant than specific labels.

$$AP(D)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1_m^{\top}y_i}\sum_{l:y_i^l}\frac{prec_i(l)}{rank_i(l)}, \tag{34}$$

where $prec_i(l)=\sum_{l:y_i^l=1}\delta(rank_i(l)\geq rank_i(l'))$, and $AP(D)\in[0,1]$.

- **Ranking loss** is the proportion of label pairings that are in reverse order, or when unrelated labels are more important than related labels.

$$RL(D)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1_m^{\top}y_i1_m^{\top}\overline{y_i}}\sum_{l:y_i^l=1}\sum_{l':y_i^l=0}(\delta(rank_i(l)\geq rank_i(l'))), \tag{35}$$

where $\overline{y_i}$ is the complement of y_i in Y, and $RL(D)\in[0,1]$.

- **Hamming loss** determines the percentage of labels that are incorrectly labeled, meaning that either a label that belongs to the instance or one that does not predicted.

$$HL(D)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\|h(x_i)\Delta y_i\|_1. \tag{36}$$

The Δ is used to denote the symmetric difference between the two sets, which is the set of values that appear exclusively in either one of the two sets.

- **Coverage Error** indicates the number of steps required to cover all of the positive labels associated with the cases by moving down the anticipated label ranking.

$$CV(D)=\frac{1}{n}\sum_{i=1}^{n}\arg\max_{l:y_i^l=1}rank_i(l)-1. \tag{37}$$

A smaller value indicates better classification performance in terms of Ranking loss, Hamming loss, and Coverage Error when the ideal value is 0. While a higher value reflects better classification performance in terms of Micro-F1, Macro-F1, and Average accuracy, the ideal value is 1.

4.3. Compared methods

We evaluate our method by comparing it to the latest multi-label feature selection methods. Each method is briefly described in the following

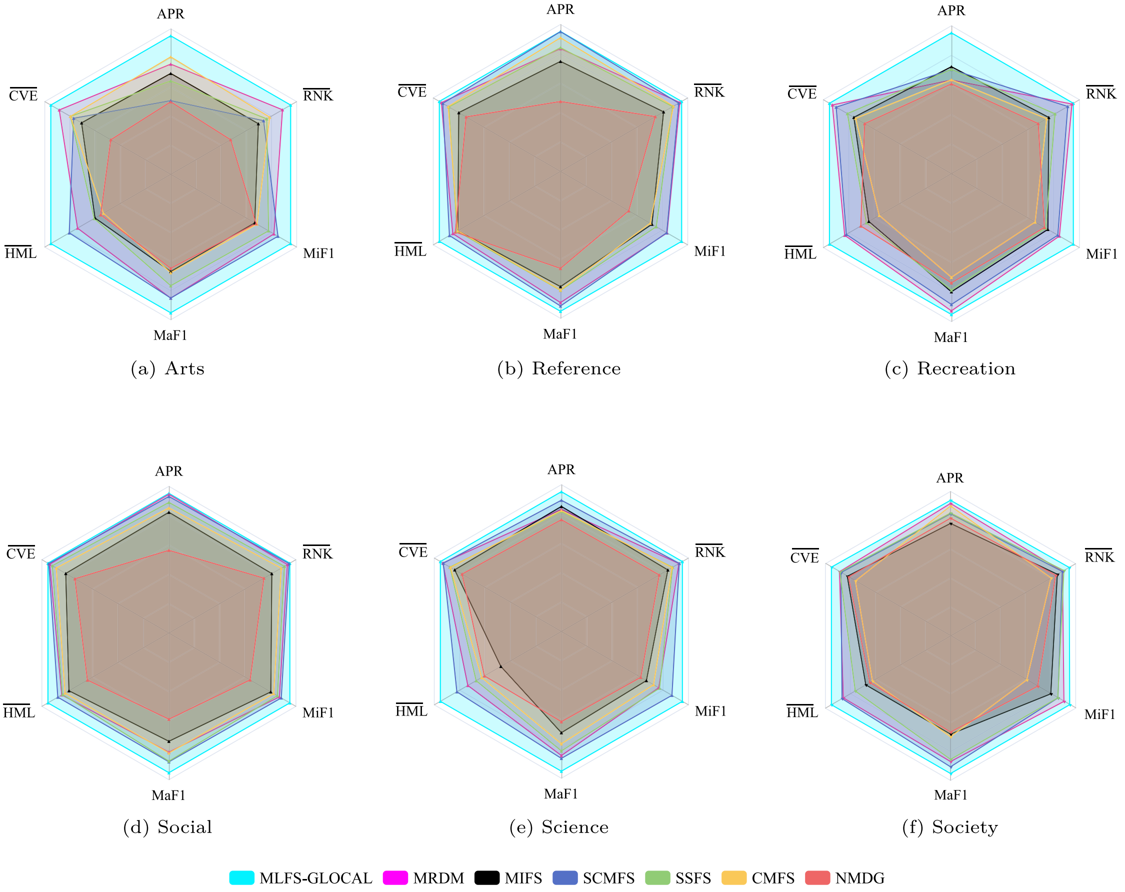
- **MDMR** (Lin et al., 2015) is a MLFS method based on information theory that chooses features by simultaneously maximizing the dependency and minimizing redundancy.
- **SCLS** (Lee & Kim, 2017) is another MLFS algorithm that uses information theory and evaluates conditional relevance using a scalable relevance assessment criterion.
- **LRFS** (Zhang, Liu, et al., 2019) is based on label redundancy. The conditional mutual information is used to build a new feature relevance term that evaluates feature information.
- **MIFS** (Jian et al., 2016) is a well-known MLFS method that makes use of latent semantics of the multi-labels to choose the most important features.
- **CMFS** (Braytee et al., 2017) is a method for studying structured information that relies on feature-side and label-side correlations.
- **SCMFS** (Hu et al., 2020) employs coupled nonnegative matrix factorization to create the shared common model.
- **SSFS** (Gao et al., 2023) presents a Latent Structure Shared (LSS) term that shares and maintains both the latent feature and label structures.
- **MRDM** (Huang & Wu, 2021) utilizes HSIC as a criterion To increase the reliance between the Manifold embedding and the class labels.
- **NMDG** (Zhang & Ma, 2022) uses the dynamic graph laplacian matrix constructed by pseudo-label in the feature selection process.

4.4. Experimental results

In this experiment, we selected the top 20% of features to determine the average performances for each technique. Tables 2–7 present the results of these experiments on the six different evaluation criteria. The best results for each dataset are indicated by bold fonts, where the higher the values, the better the classification performance. To provide a more robust evaluation of performance, each method is run 10 times and the average results are reported for all datasets. The tables demonstrate that, across most datasets, the proposed model produces the best results. In addition, the best Micro-F1 score was obtained by MLFS-GLOCAL, which had a significant lead over the second-best method on the Arts, Corel5k, Entertainment, Health, Recreation, and Social datasets. These tables show that the proposed model is first in 76 of the 84 comparison cases and comes in second in the remaining ones. These results confirm that our method can be applied to a broad spectrum of datasets, as opposed to other techniques. On average, we observed significant improvements in the values of these metrics, with 0.0379 for Micro-F1, 0.0176 for Macro-F1, 0.0088 for Average Precision, 0.1017 for Coverage Error, 0.0019 for Hamming Loss, and 0.0034 for Ranking Loss.

In addition, we use radar diagrams to show the universality of our method on six different evaluation metrics in comparison with other methods. These metrics measure various aspects of the quality and effectiveness of methods (Seyedi, Ghodsi, Akhlaghian, Jalili, & Moradi, 2019). It is worth mentioning that the Coverage error, Ranking loss, and hamming loss metrics are utilized inversely in order to maintain consistency with the other measures, so that a larger scale implies better performance for all measures. Also, to make the comparisons fair and clear, we normalize the data in Fig. 2 so that all the values are between 0.5 and 1. The more area a method covers in the radar charts, the better it is in all the evaluation metrics. Fig. 2 shows that our method has a larger area than the other methods, which means that it is more universal and superior in performance and evaluation criteria. This demonstrates that our method can handle different types of problems and situations better than the existing methods.

In assessing the efficacy of MLFS-GLOCAL and other comparative approaches in feature selection for identifying top-relevance features, we analyze their performances with different numbers of selected features. This analysis is illustrated visually using four datasets: Arts, Business, Corel5k, and Entertainment. The y-axis in Figs. 3–6 depicts the performance of various evaluation criteria, while the x-axis shows the percentage of selected features from 1% to 20%. We can compare the performance of each technique on the same dataset with the same metric. It is clear that the methods generally perform better as more features are selected. Notably, our method demonstrates superior



(a) Arts (b) Reference (c) Recreation

(d) Social (e) Science (f) Society

Fig. 2. Universality analysis of the proposed method on the evaluation criteria for six different multi-label datasets is shown by the radar chart, where the APR = Average Precision, RNK = Ranking loss, MiF1 = Micro-F1, MaF1 = Macro-F1, HML = Hamming loss, and CVE = Coverage error.

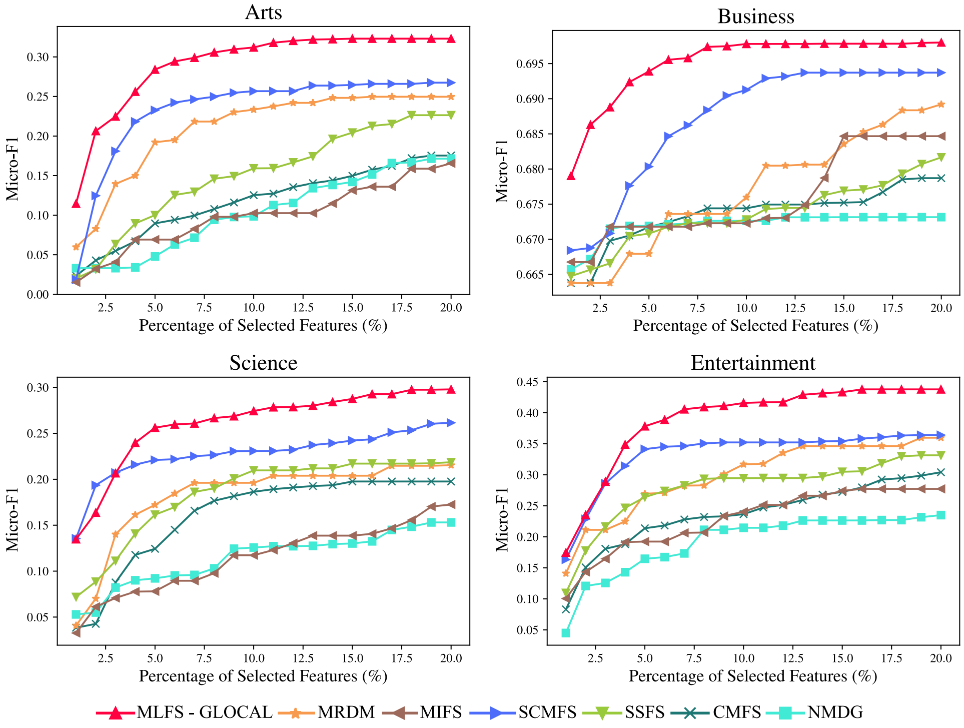


Fig. 3. Robustness analysis on four benchmark datasets in terms of Micro-F1.

Table 2
Micro-F1 results on real-world datasets. The best result is highlighted in **bold** style, while underline style indicates the second-best.

Datasets	SSFS	SCMFS	MIFS	CMFS	MRDM	NMDG	SCLS	LRFS	MDMR	MLFS-GLOCAL
Arts	0.2263	0.2674	0.1655	0.1753	0.2496	0.1714	0.1258	0.0601	0.1418	**0.3230**
Birds	0.0810	0.1757	0.0850	0.0773	0.0350	0.1300	0.0740	0.0469	0.0125	**0.2282**
Business	0.6816	0.6927	0.6846	0.6786	0.6892	0.6728	0.6760	0.6698	0.6704	**0.6966**
Computers	0.4264	0.4254	0.0388	0.4105	0.4217	0.4064	0.4310	0.4088	0.4083	**0.4511**
Corel5k	0.0276	0.0397	0.0388	0.0338	0.0392	0.0361	0.0377	0.0141	0.0208	**0.0495**
Education	0.3061	0.3131	0.2095	0.2871	0.2306	0.1191	0.1383	0.0910	0.1591	**0.3594**
Entertainment	0.3314	0.3640	0.2796	0.3041	0.3597	0.2350	0.2665	0.1219	0.2828	**0.4377**
Health	0.4887	0.5158	0.4647	0.4783	0.4904	0.4649	0.4385	0.3671	0.4193	**0.5744**
Recreation	0.2158	0.2689	0.2252	0.1694	0.2760	0.2104	0.1432	0.0693	0.1766	**0.3371**
Reference	0.4371	0.4563	0.4291	0.4259	0.4570	0.3864	0.4083	0.3752	0.3722	**0.4838**
Scene	0.5766	0.6412	0.5999	0.5229	0.5132	0.5166	0.5832	0.2789	0.5684	**0.6789**
Science	0.2185	0.2615	0.1725	0.1975	0.2153	0.1530	0.1422	0.0871	0.1406	**0.2978**
Social	0.5306	0.5485	0.4995	0.5132	0.5398	0.4035	0.4520	0.3069	0.4430	**0.5883**
Society	0.3534	0.3536	0.3429	0.3234	0.3625	0.3071	0.2904	0.2862	0.2823	**0.3711**

Table 3
Macro-F1 results on real-world datasets. The best result is highlighted in **bold** style, while underline style indicates the second-best.

Datasets	SSFS	SCMFS	MIFS	CMFS	MRDM	NMDG	SCLS	LRFS	MDMR	MLFS-GLOCAL
Arts	0.1052	0.1299	0.0765	0.0790	0.1301	0.0702	0.0534	0.0203	0.0638	**0.1597**
Birds	0.0276	0.0970	0.0356	0.0367	0.0128	0.0665	0.0228	0.0148	0.0042	**0.1172**
Business	0.0789	0.1066	0.0984	0.0762	0.0951	0.0633	0.0679	0.0427	0.0527	**0.1224**
Computers	0.1184	0.1205	0.0747	0.1128	0.1401	0.0514	0.0981	0.0503	0.0830	**0.1659**
Corel5k	0.0022	0.0032	0.0027	0.0024	0.0023	0.0019	0.0035	0.0018	0.0028	**0.0044**
Education	0.0936	0.1151	0.0657	0.0970	0.0880	0.0365	0.0488	0.0285	0.0451	**0.1228**
Entertainment	0.1736	0.1918	0.1291	0.1488	0.1907	0.1170	0.1191	0.0399	0.1138	**0.2155**
Health	0.1908	0.1991	0.1599	0.1844	0.1967	0.1405	0.1321	0.0673	0.1030	**0.2283**
Recreation	0.1317	0.1576	0.1347	0.1087	0.1691	0.1188	0.0802	0.0490	0.0905	**0.1756**
Reference	0.0940	0.1136	0.0892	0.0935	0.1093	0.0663	0.0690	0.0322	0.0613	**0.1203**
Scene	0.5719	0.6446	0.5984	0.5146	0.5044	0.5103	0.5880	0.2661	0.5588	**0.6823**
Science	0.0864	0.0947	0.0660	0.0784	0.0915	0.0538	0.0531	0.0310	0.0495	**0.1090**
Social	0.1349	0.1362	0.0974	0.1195	0.1177	0.0558	0.0937	0.0250	0.0611	**0.1567**
Society	0.1024	0.1114	0.0718	0.0693	0.1050	0.0750	0.0415	0.0362	0.0432	**0.1195**

Table 4
Average Precision results on real-world datasets. The best result is highlighted in **bold** style, while underline style indicates the second-best.

Datasets	SSFS	SCMFS	MIFS	CMFS	MRDM	NMDG	SCLS	LRFS	MDMR	MLFS-GLOCAL
Arts	0.3725	0.3636	0.3759	0.3834	0.3801	0.3631	0.3641	0.3676	0.3616	**0.3929**
Birds	0.1783	0.1853	0.2149	0.1918	0.1778	**0.2420**	0.1818	0.1782	0.1813	0.2237
Business	0.8311	0.8301	0.8280	0.8324	0.8382	0.8249	0.8308	0.8275	0.8142	**0.8407**
Computers	0.5644	0.5821	0.5727	0.5666	0.5632	0.5650	**0.5868**	0.5596	0.5798	0.5838
Corel5K	0.1589	0.1623	0.1554	0.1575	0.157	0.1545	0.1411	0.1484	0.1582	**0.1653**
Education	0.4195	0.4155	0.4193	0.4276	0.4298	0.4236	0.4225	0.4168	0.3815	**0.4394**
Entertainment	0.4467	0.4527	0.4469	0.4510	0.4525	0.4373	0.4306	0.4203	0.4199	**0.4568**
Health	0.5992	0.5894	0.6005	0.5912	0.5956	0.5833	0.5632	0.5411	0.5530	**0.6045**
Recreation	0.3523	0.3521	0.3553	0.3405	0.3402	0.3359	0.3258	0.3362	0.3139	**0.3936**
Reference	0.5544	0.5636	0.5465	0.5601	0.5538	0.5235	0.5312	0.5235	0.5300	**0.5637**
Scene	0.6681	0.7044	0.6866	0.6412	0.6217	0.5825	0.6876	0.5009	0.6951	**0.7482**
Science	0.3461	0.3553	0.3502	0.3469	0.3478	0.3394	0.3224	0.3379	0.3050	**0.3623**
Social	0.6352	0.6466	0.6162	0.6246	0.6508	0.5431	0.5819	0.5183	0.5770	**0.6526**
Society	0.4952	0.4949	0.4915	0.4934	0.4986	0.4977	0.4760	**0.5004**	0.4871	0.4997

Table 5
Ranking Loss results on real-world datasets. The best result is highlighted in **bold** style, while underline style indicates the second-best.

Datasets	SSFS	SCMFS	MIFS	CMFS	MRDM	NMDG	SCLS	LRFS	MDMR	MLFS-GLOCAL
Arts	0.2066	0.2077	0.2095	0.2056	0.2012	0.2190	0.1998	0.2117	0.2124	**0.1984**
Birds	0.1699	0.1571	0.1437	0.1634	0.1781	**0.1092**	0.1756	0.1694	0.1579	0.1335
Business	0.0528	**0.0488**	0.0522	0.0526	0.0494	0.0547	0.0580	0.0583	0.0592	0.0490
Computers	0.1207	0.1172	0.1218	0.1208	0.1194	0.1254	0.1250	0.1322	0.1226	**0.1148**
Corel5k	0.2085	0.2078	0.2171	0.2080	0.2088	0.2162	0.2172	0.2146	0.2161	**0.2063**
Education	0.1211	0.1216	0.1274	0.1201	0.1231	0.1356	0.1353	0.1313	0.1374	**0.1168**
Entertainment	0.1664	0.1609	0.1695	0.1685	0.1641	0.1697	0.1851	0.1874	0.1726	**0.1577**
Health	0.0804	0.0807	0.0836	0.0803	0.0790	0.0895	0.0979	0.1084	0.1025	**0.0789**
Recreation	0.2409	0.2325	0.2451	0.2465	0.2297	0.2521	0.2693	0.2605	0.2575	**0.2288**
Reference	0.1009	0.0986	0.1043	0.1006	0.0982	0.1074	0.1116	0.1203	0.1176	**0.0975**
Scene	0.2028	0.1780	0.1890	0.2280	0.2418	0.2659	0.1873	0.3723	0.1810	**0.1473**
Science	0.1635	0.1583	0.1672	0.1639	0.1586	0.1739	0.2030	0.2015	0.2016	**0.1559**
Social	0.0757	0.0732	0.0825	0.0774	0.0741	0.0867	0.0863	0.1053	0.0957	**0.0728**
Society	0.1800	0.1813	0.1852	0.1865	0.1811	0.1902	**0.1754**	0.2058	0.2226	0.1758

Table 6
Hamming Loss results on real-world datasets. The best result is highlighted in **bold** style, while underline style indicates the second-best.

Datasets	SSFS	SCMFS	MIFS	CMFS	MRDM	NMDG	SCLS	LRFS	MDMR	MLFS-GLOCAL
Arts	0.0615	0.0588	0.0616	0.0625	0.0597	0.0622	0.0631	0.0633	0.0627	**0.0568**
Birds	0.0554	0.0527	0.0532	0.0532	0.0537	0.0539	0.0529	0.0529	0.0514	**0.0497**
Business	0.0288	0.0278	0.0281	0.0290	0.0286	0.0287	0.0286	0.0288	0.0288	**0.0276**
Computers	0.0393	0.0394	0.0404	0.0402	0.3979	0.0411	0.0396	0.0422	0.0412	**0.0388**
Corel5k	0.0095170	0.009506	0.009495	0.009506	0.009504	0.009506	0.009516	0.009512	0.009535	**0.009495**
Education	0.0414	0.0410	0.0437	0.0418	0.0436	0.0450	0.0443	0.0444	0.0442	**0.0392**
Entertainment	0.0615	0.0594	0.0641	0.0630	0.0613	0.0658	0.0634	0.0676	0.0629	**0.0555**
Health	0.0427	0.0407	0.0436	0.0433	0.0423	0.0430	0.0461	0.0506	0.0466	**0.0382**
Recreation	0.0617	0.0592	0.0620	0.0633	0.0590	0.0610	0.0644	0.0647	0.0628	**0.0571**
Reference	0.0296	0.0287	0.0297	0.0297	0.0291	0.0294	0.0323	0.0347	0.0322	**0.0275**
Scene	0.1333	0.1155	0.1249	0.1494	0.1509	0.1499	0.1471	0.1879	0.1343	**0.1074**
Science	0.0349	0.0342	0.0358	0.0351	0.0346	0.0352	0.0357	0.0358	0.0354	**0.0336**
Social	0.0243	0.0236	0.0253	0.0247	0.0242	0.0281	0.0275	0.0313	0.0262	**0.0221**
Society	0.0563	0.0553	0.0571	0.0574	0.0554	0.0576	0.0590	0.0590	0.0583	**0.0545**

Table 7
Coverage Error results on real-world datasets. The best result is highlighted in **bold** style, while underline style indicates the second-best.

Datasets	SSFS	SCMFS	MIFS	CMFS	MRDM	NMDG	SCLS	LRFS	MDMR	MLFS-GLOCAL
Arts	7.820	7.852	7.930	7.820	7.710	8.217	7.662	8.045	8.059	**7.627**
Birds	4.764	4.367	4.154	4.588	4.910	**3.443**	4.792	4.752	4.427	3.941
Business	3.731	**3.584**	3.686	3.740	3.609	3.808	3.933	4.009	4.060	3.588
Computers	6.440	6.340	6.458	6.495	6.324	6.619	6.723	6.849	6.673	**6.237**
Corel5k	164.89	164.36	170.95	164.83	164.81	172.02	171.95	171.96	172.00	**163.63**
Education	5.870	5.882	6.093	5.820	6.019	6.453	6.505	6.412	6.621	**5.735**
Entertainment	5.190	5.075	5.269	5.254	5.149	5.027	5.610	5.694	5.359	**5.010**
Health	4.985	4.955	5.096	4.960	4.932	5.378	5.663	6.032	5.920	**4.903**
Recreation	7.125	6.944	7.231	7.259	6.882	7.406	7.824	7.678	7.511	**6.835**
Reference	4.820	4.742	4.946	4.811	4.724	5.039	5.194	5.470	5.411	**4.698**
Scene	2.116	1.990	2.049	2.247	2.311	2.434	2.043	2.967	2.011	**1.837**
Science	8.867	8.630	9.040	8.885	8.613	9.303	10.698	10.637	10.624	**8.525**
Social	4.820	4.739	5.131	4.911	4.761	5.341	5.339	6.116	5.841	**4.719**
Society	7.621	7.640	7.754	7.775	7.613	7.899	7.577	8.536	7.580	**7.470**

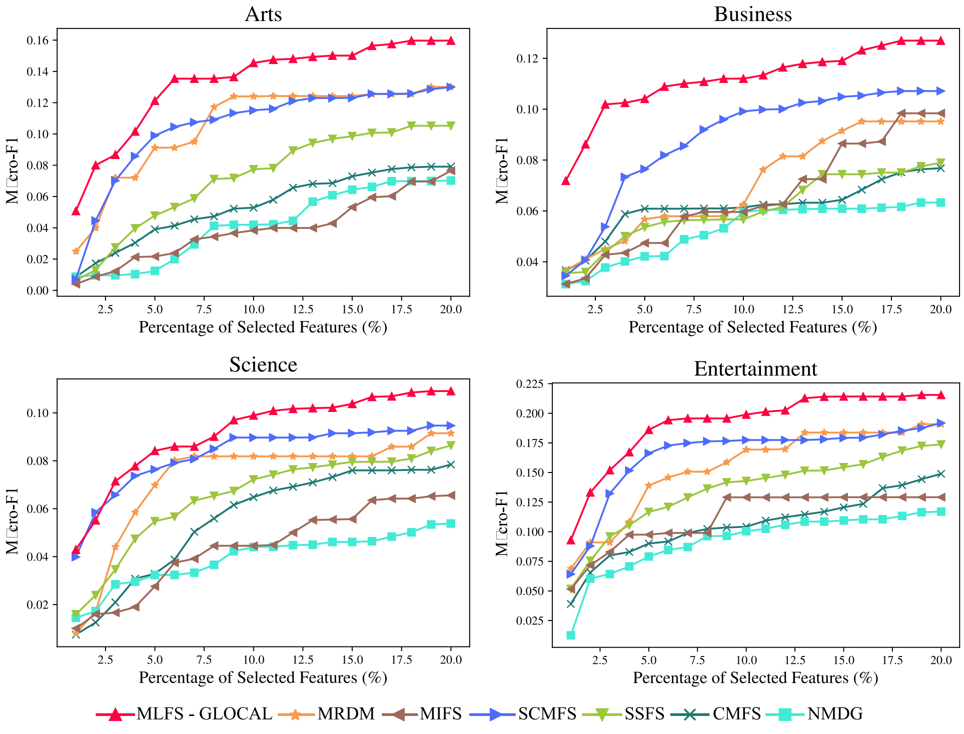


Fig. 4. Robustness analysis on four benchmark datasets in terms of Macro-F1.

performance with a small number of features, as evident in Figs. 3–6, highlighting the efficacy of our proposed model in selecting more relevant features. Specifically, Figs. 3–5 show the results for Arts, Business, Corel5k, and Entertainment datasets, where the Micro-F1, Macro-F1, and hamming loss metrics of our MLFS-GLOCAL model are markedly better than those of several other algorithms.

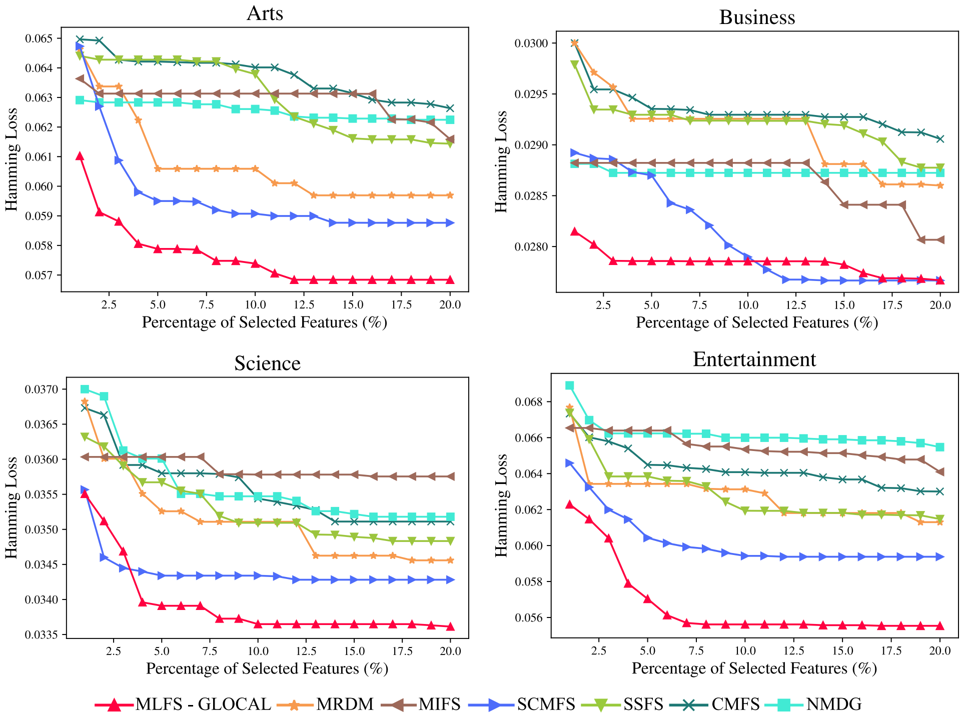


Fig. 5. Robustness analysis on four benchmark datasets in terms of Hamming Loss.

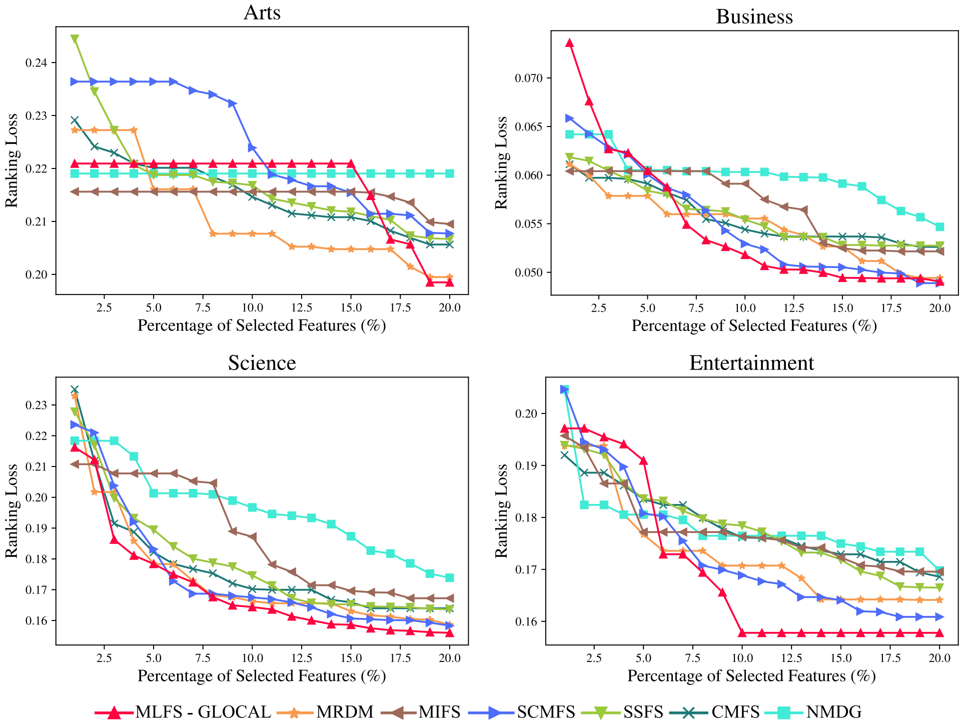


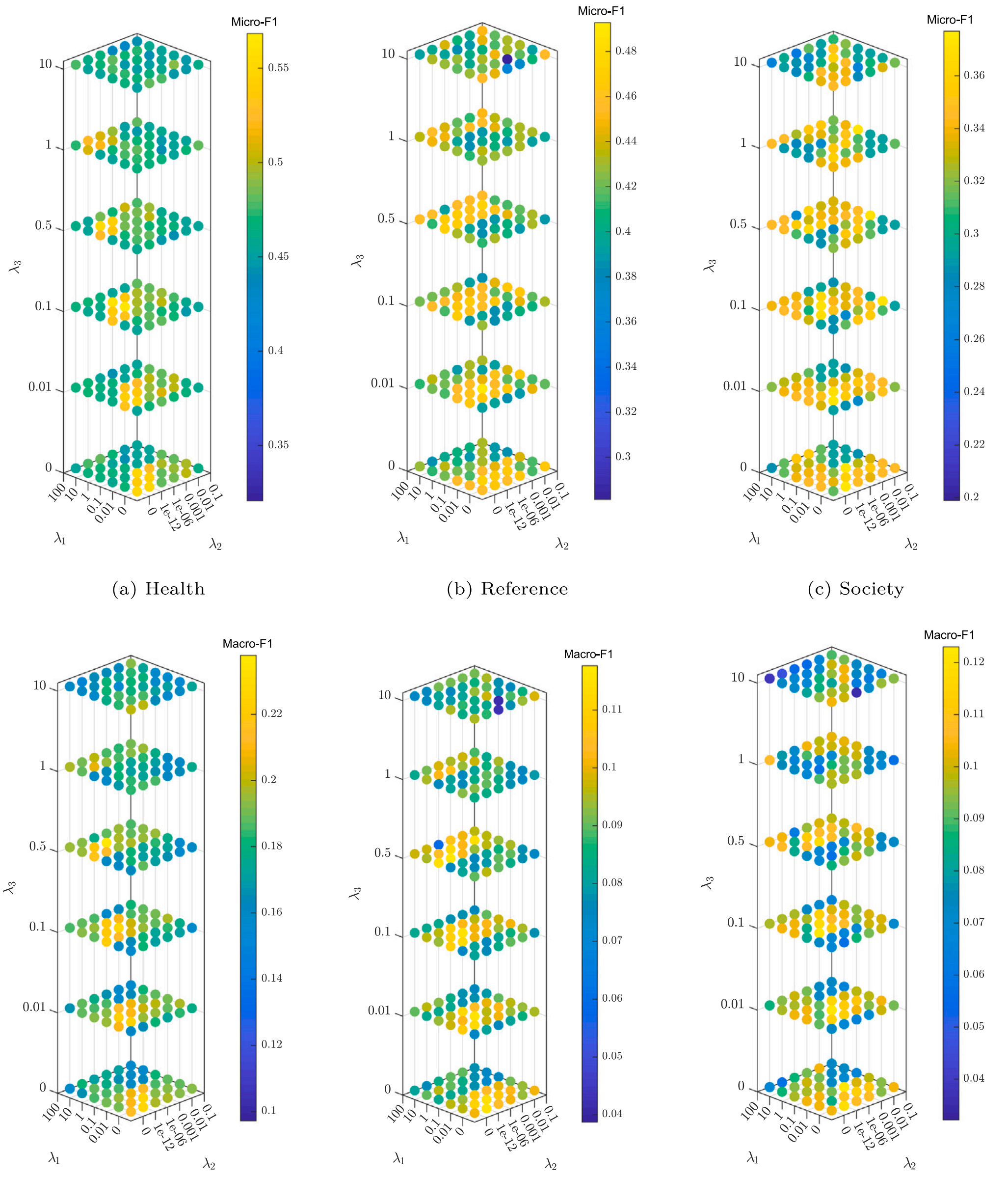
Fig. 6. Robustness analysis on four benchmark datasets in terms of Ranking Loss.

4.5. Analyzing the sensitivity of parameters

In this section, we analyzed the influence of the hyperparameters, including the graph regularization parameter λ_1, the Global and Local label correlation parameter λ_2, and the sparsity parameter λ_3. Figs. 7 and 8 illustrates the Micro-F1, Macro-F1, Hamming Loss, and Ranking Loss metrics of our method with various λ_1, λ_2, and λ_3 on the six datasets. These Figures are plotted in 3D, i.e., three axes relate to λ_1, λ_2, and λ_3.

4.5.1. Choosing the value of parameter λ_1

This parameter controls the manifold regularization term's effectiveness in the proposed method. In this parameter sensitivity analysis, the values in the set $\{0, 0.01, 0.1, 1, 10, 100\}$ are selected for λ_1 parameter



(a) Health (b) Reference (c) Society

(d) Health (e) Reference (f) Society

Fig. 7. Parameter Analysis of our method with respect to the parameters λ_1, λ_2 and λ_3 on six real-world datasets in terms of Micro-F1 and Macro-F1.

across all datasets. From Figs. 7 and 8, it can be deduced that the optimal value for this parameter is typically less than 1.

4.5.2. Choosing the value of parameter λ_2

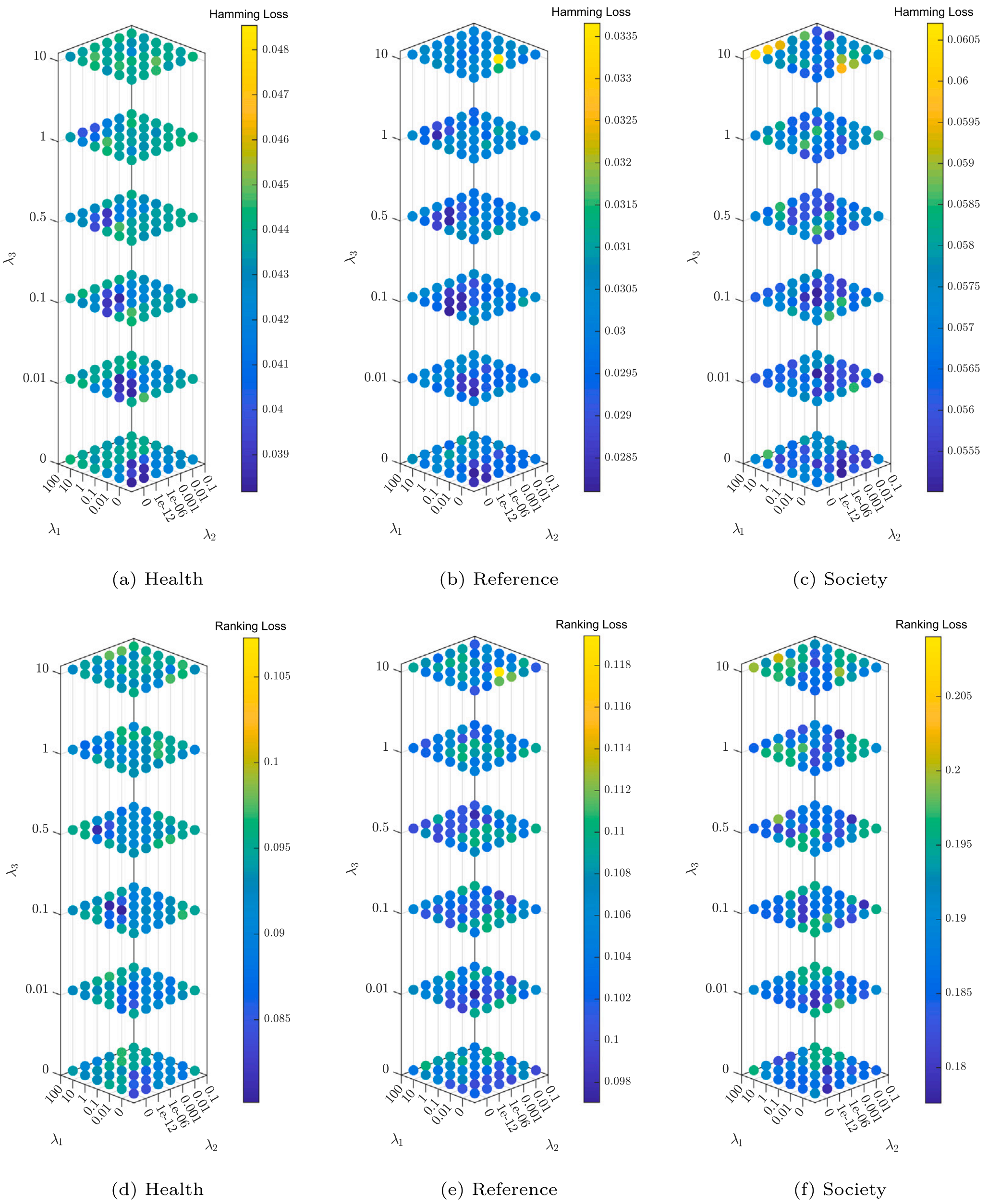
This parameter shows the effectiveness rate on global and local label correlation in our method where the analyzed values for λ_2 parameter are $\{0, 10^{-12}, 10^{-6}, 10^{-3}, 10^{-2}, 10^{-1}\}$. As we can observe in Figs. 7 and 8, λ_2 with small values usually results a better performance in terms of Micro-F1 and Macro-F1 and with high values usually has better in terms of Hamming Loss and Coverage Error measures on the four datasets. Values near to zero or very large ones for this parameter may not perform relatively well.

4.5.3. Choosing the value of parameter λ_3

The sparsity regularization term in the MLFS-GLOCAL method is controlled by λ_3. The range of λ_3 is $\{0, 0.001, 0.1, 0.5, 1, 10\}$. The results show that λ_3 is a delicate quantity that typically requires careful adjusting. The results indicate that choosing values less than 1 for this parameter is usually preferable.

4.5.4. Varying the latent representation dimensionality k

Fig. 9 analyzes the effect of varying the matrix latent space k on the Arts, Business, Education, Entertainment, Science, Scene, and Birds datasets in terms of Micro-F1, Macro-F1, Hamming Loss, and Ranking Loss evaluation metrics. As depicted in Fig. 9, generally, the model's performance shows low sensitivity to changes in the parameter k. However, some changes are observed in the results for the Scene, Birds, and Entertainment datasets when the value of k is altered. On the Scene dataset, particularly in terms of Ranking Loss, as well as on the Birds dataset in terms of Macro-F1, the k parameter with a small value cannot capture enough information. In these cases, performance improves with an increase in k. However, when k becomes excessively large, the low-rank structure is underutilized, leading to a decline in performance. Furthermore, on the Birds dataset, we observe sensitivity



(a) Health (b) Reference (c) Society

(d) Health (e) Reference (f) Society

Fig. 8. Parameter Analysis of our method with respect to the parameters λ_1, λ_2 and λ_3 on six real-world datasets in terms of Hamming Loss and Ranking Loss.

to smaller values of k, specifically in terms of Micro-F1 and Macro-F1. A smaller k appears to have an impact on performance in this dataset. As k increases, the model's performance improves. Nonetheless, when k becomes excessively large, the low-rank structure is not fully leveraged, resulting in a decline in performance. Therefore, while k generally exhibits low sensitivity, its proper selection should be considered in achieving optimal results across different datasets.

4.6. Ablation study

To assess the influence of different components in the proposed method (i.e., MLFS-GLOCAL), we performed ablation experiments on six datasets (Arts, Health, Education, Entertainment, Science, and Birds). The employed evaluation metrics include Micro-F1, Ranking Loss, and Coverage Error. Within this analysis, we specified five scenarios to illustrate the efficacy of each component within our model. In Fig. 10, Case I ($\lambda_1 = \lambda_2 = \lambda_3 = 0$) represents the basic model without any regularization terms (14). In Case II ($\lambda_2 = \lambda_3 = 0$), we analyzed the model with the local structure regularization. Case III ($\lambda_1 = \lambda_3 = 0$) indicates the model only with global and local label correlation. Case IV($\lambda_1 = \lambda_2 = 0$) refers to the model with the $\ell_{2,1}$ sparse regularization. Finally, case V represents the final model (i.e., MLFS-GLOCAL) with all components. The results reveal that Case II, incorporating general regularization in representation models like embedded feature selection methods, improves the basic model. However, introducing global and local label correlation (Case III), a specialized regularization for multi-label learning, yields even greater enhancement than Case II. Additionally, Case IV underscores the importance of feature weight sparsity in the feature selection task. In summary, the ablation experiments demonstrate the incremental contributions of each regularization term to the overall performance of the model (Case V). The findings highlight the importance of considering both general and specialized

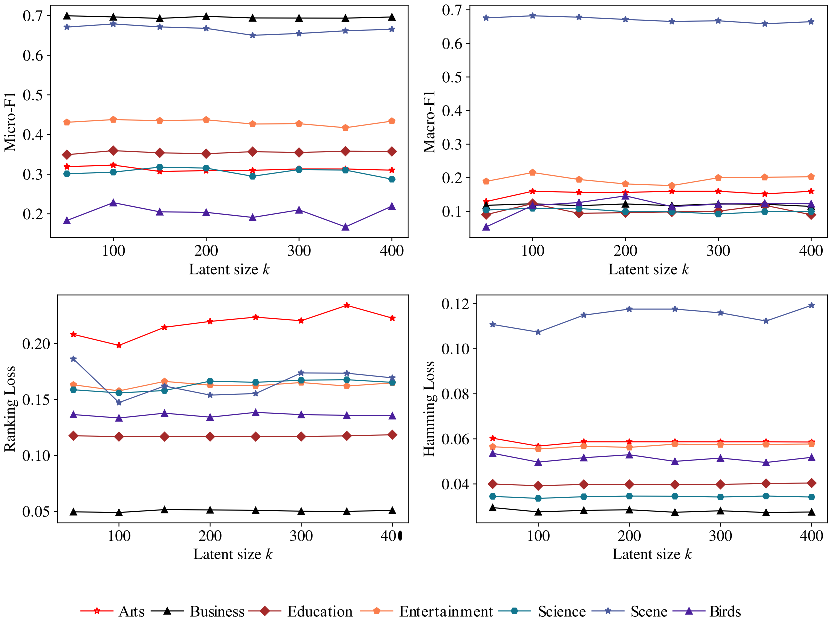


Fig. 9. Adjusting the dimensionality of latent representations for Arts, Business, Education, Entertainment, Science, Scene, and Birds datasets.

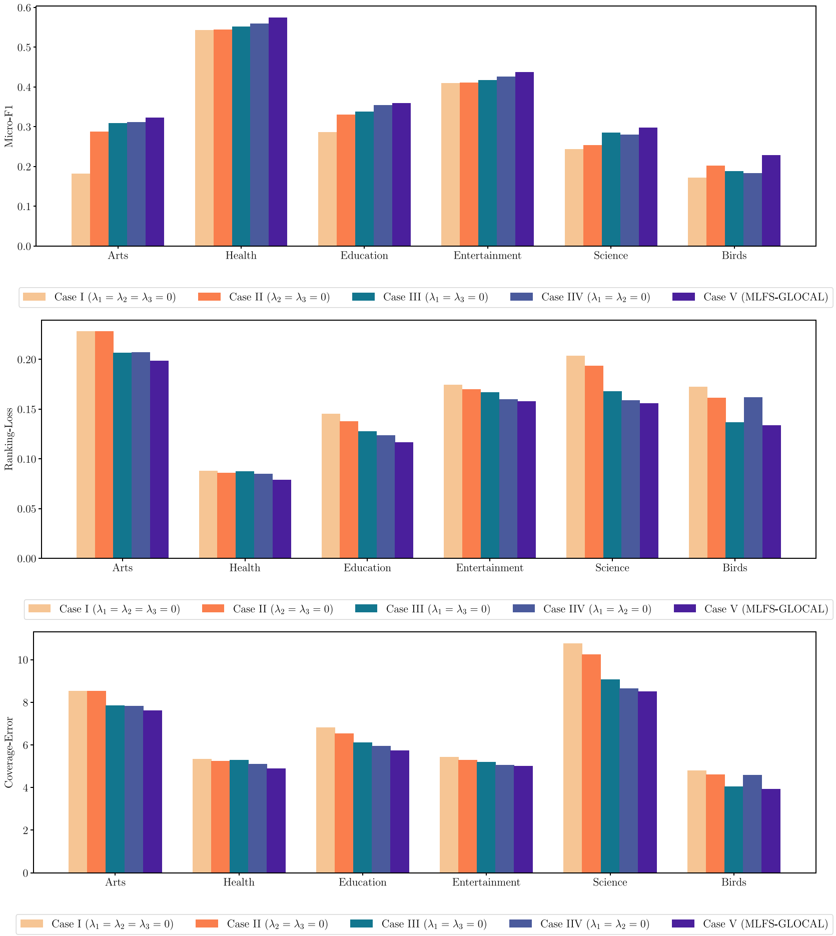
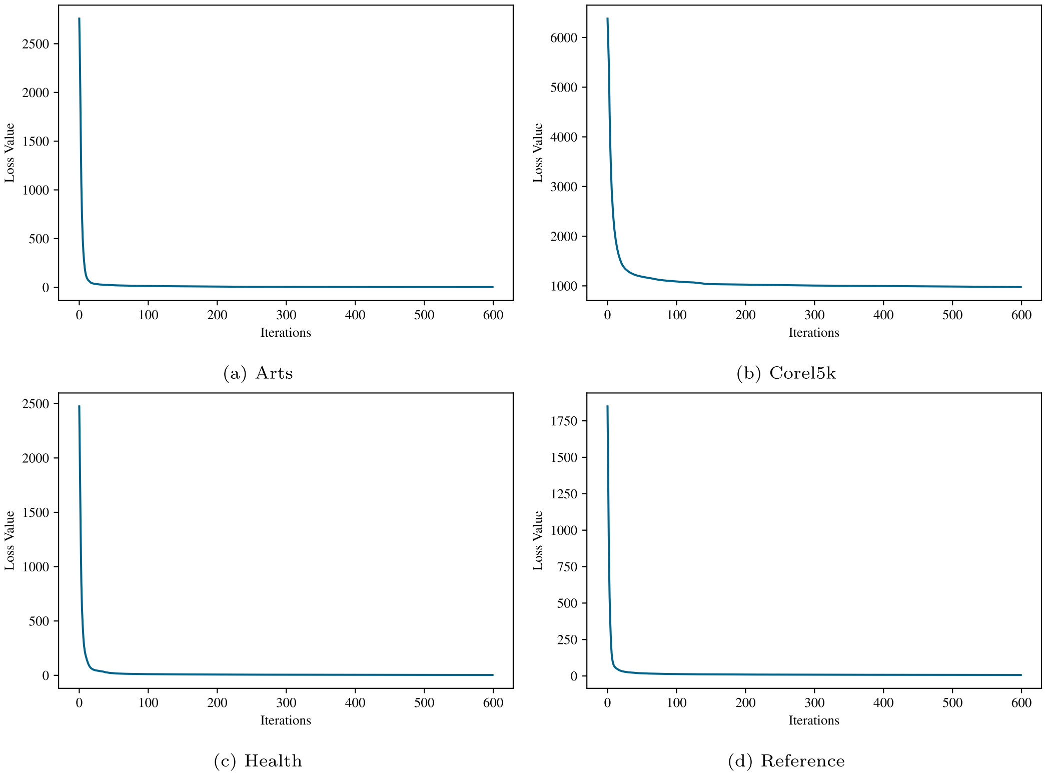


Fig. 10. Ablation study of the method on six real-world datasets in terms of Micro-F1, Ranking Loss, and Coverage Error, where Case I is basic model without any regularization terms, Case II is basic model with local structure regularization, Case III is basic model with global and local label correlation, Case IV is basic model with sparse regularization, and Case V is final model (MLFS-GLOCAL) with all components.



(a) Arts

(b) Corel5k

(c) Health

(d) Reference

Fig. 11. Convergence analysis of MLFS-GLOCAL model on the Arts, Corel5k, Health, and Reference datasets.

regularization techniques in multi-label feature selection, as well as the synergistic effect of combining these regularization terms in the proposed MLFS-GLOCAL model.

4.7. Convergence

In this section, we evaluate the convergence behavior of our proposed MLFS-GLOCAL model (22) by conducting experiments on four datasets: Arts, Corel5k, Health, and Reference. We apply Algorithm 1 to each dataset and run it for 600 iterations. We plot the objective function value against the number of iterations in Fig. 11 to show how our method converges to a solution. As can be seen from Fig. 11, the objective function value decreases rapidly and steadily in the initial iterations, indicating that our method quickly approaches an optimal solution. In the later iterations, the objective function value changes very slightly, suggesting that our method has reached a near-optimal solution. This demonstrates that our method has fast and stable convergence performance.

5. Conclusion

In this study, we propose a novel multi-label feature selection method by considering implicit and explicit label correlation. The global label correlation helps to exploit the underlying structure of the label space. It allows the model to learn relationships and dependencies between labels. Local label correlation, on the other hand, refers to the label associations that are specific to a local context. By considering the correlation between the labels and incorporating them into the feature representation, the model identifies the features that are most relevant to the labels. In addition, MLFS-GLOCAL learns the shared common mode between the feature matrix and label matrix to extract implicit label correlation and guide feature selection. This low-dimensional embedding is constrained through manifold regularization, which means that it has a similar local structure to the original data and retains the most valuable information in the data. An alternating optimization-based iterative algorithm is developed to solve the objective function with $L_{2,1}$-norm regularization. Finally, extensive experiments over variable multi-label datasets demonstrate the effectiveness of the proposed MLFS-GLOCAL against some state-of-the-art feature selection methods.

One limitation of our study is that the calculated global and local label correlations may suffer from noise due to labels with few positive instances. Future research should focus on learning label correlation methods to improve model performance. Also, due to the expensive label information in the real world, we can focus on semi-supervised (Chavoshinejad, Seyedi, Akhlaghian Tab, & Salahian, 2023; Seyedi, Moradi, & Tab, 2017) or missing multi-label learning methods (Mahmoodi, Seyedi, Akhlaghian Tab, & Abdollahpouri, 2023; Seyedi, Akhlaghian Tab, Lotfi, Salahian, & Chavoshinejad, 2023) based on label correlations. It is also interesting to extend our work to a multi-label multi-view setting, where the data is represented by multiple feature sets or views, and each instance is associated with multiple labels.

CRediT authorship contribution statement

Mohammad Faraji: Methodology, Implementation, Writing – original draft, Visualization. **Seyed Amjad Seyedi:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Fardin Akhlaghian Tab:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Reza Mahmoodi:** Writing – review & editing, Visualization, Validation.

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.

Data availability

Data will be made available on request.

References

Abdollahi, R., Amjad Seyedi, S., & Reza Noorimehr, M. (2020). Asymmetric semi-nonnegative matrix factorization for directed graph clustering. In *2020 10th International conference on computer and knowledge engineering* ICCKE, (pp. 323–328).

Braytee, A., Liu, W., Catchpoole, D. R., & Kennedy, P. J. (2017). Multi-label feature selection using correlation information. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 1649–1656).

Cai, D., He, X., Han, J., & Huang, T. S. (2010). Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(8), 1548–1560.

Cai, X., Nie, F., & Huang, H. (2013). Exact top-k feature selection via L2,0-norm constraint. In *PInternational joint conference on artificial intelligence* (pp. 1240–1246).

Chavoshinejad, J., Seyedi, S. A., Akhlaghian Tab, F., & Salahian, N. (2023). Self-supervised semi-supervised nonnegative matrix factorization for data clustering. *Pattern Recognition*, *137*, Article 109282.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, *3*(1), 140.

Doquire, G., & Verleysen, M. (2013). Mutual information-based feature selection for multilabel classification. *Neurocomputing*, *122*, 148–155.

Fan, Y., Liu, J., Liu, P., Du, Y., Lan, W., & Wu, S. (2021). Manifold learning with structured subspace for multi-label feature selection. *Pattern Recognition*, *120*, Article 108169.

Fan, Y., Liu, J., Weng, W., Chen, B., Chen, Y., & Wu, S. (2021). Multi-label feature selection with local discriminant model and label correlations. *Neurocomputing*, *442*, 98–115.

Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., & Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine Learning*, *73*, 133–153.

Gao, W., Hu, L., & Zhang, P. (2018). Class-specific mutual information variation for feature selection. *Pattern Recognition*, *79*, 328–339.

Gao, W., Li, Y., & Hu, L. (2023). Multilabel feature selection with constrained latent structure shared term. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(3), 1253–1262.

Hajiveiseh, A., Seyedi, S. A., & Akhlaghian Tab, F. (2024). Deep asymmetric nonnegative matrix factorization for graph clustering. *Pattern Recognition*, *148*, Article 110179.

Hong, R., Wang, M., Gao, Y., Tao, D., Li, X., & Wu, X. (2013). Image annotation by multiple-instance learning with discriminative feature mapping and selection. *IEEE Transactions on Cybernetics*, *44*(5), 669–680.

Hu, L., Li, Y., Gao, W., Zhang, P., & Hu, J. (2020). Multi-label feature selection with shared common mode. *Pattern Recognition*, *104*, Article 107344.

Huang, J., Li, G., Huang, Q., & Wu, X. (2017). Joint feature selection and classification for multilabel learning. *IEEE Transactions on Cybernetics*, *48*(3), 876–889.

Huang, R., & Wu, Z. (2021). Multi-label feature selection via manifold regularization and dependence maximization. *Pattern Recognition*, *120*, Article 108149.

Huang, S.-J., & Zhou, Z.-H. (2012). Multi-label learning by exploiting label correlations locally. *volume 26*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 949–955).

Ji, S., Tang, L., Yu, S., & Ye, J. (2008). Extracting shared subspace for multi-label classification. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 381–389).

Jian, L., Li, J., Shu, K., & Liu, H. (2016). Multi-label informed feature selection. In *Proceedings of the Twenty-Fifth international joint conference on artificial intelligence* (pp. 1627—1633).

Kumar, S., & Rastogi, R. (2022). Low rank label subspace transformation for multi-label learning with missing labels. *Information Sciences*, *596*, 53–72.

Lee, J., & Kim, D.-W. (2017). SCLS: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognition*, *66*, 342–352.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Li, F., Miao, D., & Pedrycz, W. (2017). Granular multi-label feature selection based on mutual information. *Pattern Recognition*, *67*, 410–423.

Lin, Y., Hu, Q., Liu, J., & Duan, J. (2015). Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing*, *168*, 92–103.

Lin, Y., Hu, Q., Liu, J., Li, J., & Wu, X. (2017). Streaming feature selection for multilabel learning based on fuzzy mutual information. *IEEE Transactions on Fuzzy Systems*, *25*(6), 1491–1507.

Liu, J., Lin, Y., Li, Y., Weng, W., & Wu, S. (2018). Online multi-label streaming feature selection based on neighborhood rough set. *Pattern Recognition*, *84*, 273–287.

Liu, J., Lin, Y., Wu, S., & Wang, C. (2018). Online multi-label group feature selection. *Knowledge-Based Systems*, *143*, 42–57.

Mahmoodi, R., Seyedi, S. A., Akhlaghian Tab, F., & Abdollahpouri, A. (2023). Link prediction by adversarial nonnegative matrix factorization. *Knowledge-Based Systems*, *280*, Article 110998.

Nie, F., Huang, H., Cai, X., & Ding, C. (2010). Efficient and robust feature selection via joint l2,1-norms minimization. *volume 23*, In *Advances in neural information processing systems* (pp. 1813–1821).

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, *85*, 333–359.

Seyedi, S. A., Akhlaghian Tab, F., Lotfi, A., Salahian, N., & Chavoshinejad, J. (2023). Elastic adversarial deep nonnegative matrix factorization for matrix completion. *Information Sciences*, *621*, 562–579.

Seyedi, S. A., Ghodsi, S. S., Akhlaghian, F., Jalili, M., & Moradi, P. (2019). Self-paced multi-label learning with diversity. In *Asian conference on machine learning* (pp. 790–805). PMLR.

Seyedi, S. A., Moradi, P., & Tab, F. A. (2017). A weakly-supervised factorization method with dynamic graph embedding. In *2017 Artificial intelligence and signal processing conference* AISP, (pp. 213–218).

Shajarian, Z., Seyedi, S. A., & Moradi, P. (2017). A clustering-based matrix factorization method to improve the accuracy of recommendation systems. In *2017 Iranian conference on electrical engineering* ICEE, (pp. 2241–2246).

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550.

Tang, B., Kay, S., & He, H. (2016). Toward optimal feature selection in naive Bayes for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2508–2521.

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2011). Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing*, *2011*(1), 1–9.

Wang, P., & Domeniconi, C. (2008). Building semantic kernels for text classification using wikipedia. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 713–721).

Weng, W., Lin, Y., Wu, S., Li, Y., & Kang, Y. (2018). Multi-label learning based on label-specific features and local pairwise label correlation. *Neurocomputing*, *273*, 385–394.

Xie, J., Wang, M., Xu, S., Huang, Z., & Grant, P. W. (2021). The unsupervised feature selection algorithms based on standard deviation and cosine similarity for genomic data analysis. *Frontiers in Genetics*, *12*, Article 684100.

Zhang, P., Liu, G., & Gao, W. (2019). Distinguishing two types of labels for multi-label feature selection. *Pattern Recognition*, *95*, 72–82.

Zhang, J., Luo, Z., Li, C., Zhou, C., & Li, S. (2019). Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recognition*, *95*, 136–150.

Zhang, Y., & Ma, Y. (2022). Non-negative multi-label feature selection with dynamic graph constraints. *Knowledge-Based Systems*, *238*, Article 107924.

Zhang, M.-L., & Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048.

Zhang, M.-L., & Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1819–1837.

Zhao, D., Gao, Q., Lu, Y., & Sun, D. (2022). Learning multi-label label-specific features via global and local label correlations. *Soft Computing*, *26*(5), 2225–2239.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, *5*(1), 44–53.

Zhu, Y., Kwok, J. T., & Zhou, Z.-H. (2017). Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, *30*(6), 1081–1094.

Zhu, P., Xu, Q., Hu, Q., Zhang, C., & Zhao, H. (2018). Multi-label feature selection with missing labels. *Pattern Recognition*, *74*, 488–502.