Graph Random Walk with Feature-Label Space Alignment: A Multi-Label Feature Selection Method

Wanfu Gao[1,2] , **Jun Gao**[1,2*] , **Qingqi Han**[1,2] , **Hanlin Pan**[1,2] and **Kunpeng Liu**[3]

[1]College of Computer Science and Technology, Jilin University, China

[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China

[3]Department of Computer Science, Portland State University, Portland, OR 97201 USA

gaowf@jlu.edu.cn, gaocheng23@mails.jlu.edu.cn, hanqq22@mails.jlu.edu.cn, panhl23@mails.jlu.edu.cn, kunpeng@pdx.edu

Abstract

The rapid growth in feature dimension may introduce implicit associations between features and labels in multi-label datasets, making the relationships between features and labels increasingly complex. Moreover, existing methods often adopt low-dimensional linear decomposition to explore the associations between features and labels. However, linear decomposition struggles to capture complex nonlinear associations and may lead to misalignment between the feature space and the label space. To address these two critical challenges, we propose innovative solutions. First, we design a random walk graph that integrates feature-feature, label-label, and feature-label relationships to accurately capture nonlinear and implicit indirect associations, while optimizing the latent representations of associations between features and labels after low-rank decomposition. Second, we align the variable spaces by leveraging low-dimensional representation coefficients, while preserving the manifold structure between the original high-dimensional multi-label data and the low-dimensional representation space. Extensive experiments and ablation studies conducted on seven benchmark datasets and three representative datasets using various evaluation metrics demonstrate the superiority of the proposed method.

1 Introduction

In recent years, with the exponential growth of feature dimension, feature selection in high-dimensional datasets has become a critical task. By eliminating irrelevant and redundant features while retaining important and relevant ones, feature selection effectively reduces computational time and storage costs [Li *et al.*, 2017]. As a result, feature selection methods have been widely applied across various domains, including text mining [Jin *et al.*, 2023], image annotation [Kong *et*
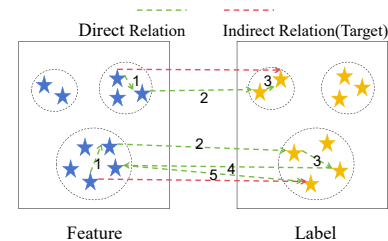
---

*Corresponding author

Figure 1: The diagram illustrates the direct and indirect associations between feature subsets and label subsets. For example, the underlying indirect relationships between features and labels can be captured through the direct relationships of 1, 2 and 3 above.

al., 2012], music classification [Silla Jr *et al.*, 2008], microarray data analysis [Bommert *et al.*, 2022], and biomarker discovery in proteomics [Tang *et al.*, 2021]. Generally, feature selection methods can be categorized into three main types: filter-based methods, wrapper-based methods, and embedded methods [Yao *et al.*, 2017].

Filter-based methods operate independently of the learning method [Zhang *et al.*, 2021]. In contrast, wrapper-based methods [Al-Yaseen *et al.*, 2022] iteratively evaluate and select subsets of features to identify the optimal feature subset. Embedded methods, on the other hand, integrate the feature selection process directly into the learning method [Yamada *et al.*, 2020]. Among embedded methods, many methods leverage sparse regularization to impose constraints that enhance the structural information of the data, thereby capturing the underlying structure more accurately and improving the generalization ability of the method. However, a key challenge for feature selection methods in multi-label learning problems is computing the relevance between features and labels.

The goal of multi-label learning is to assign multiple correct labels to a given instance [Zhu *et al.*, 2018], which poses new challenges compared to traditional single-label learning. Some studies address multi-label problems by transforming them into multiple single-label problems [Lee and Kim, 2015; Lee and Kim, 2017], but these methods often overlook the correlations between labels. Other methods handle multi-label problems by creating label sets [Wang *et al.*, 2021b],

grouping related labels together and converting the multi-label problem into a multi-class problem. However, these methods struggle to accurately capture the complex high-order correlations between features and labels in multi-label datasets.

In the real-world scenarios, a label can be associated with multiple other labels, making it inadequate to consider only pairwise label correlations [Zhang and Gao, 2021; Lim and Kim, 2020]. Consequently, the complex and deep relationships between labels have become a key focus of research [Yu *et al.*, 2021; Li *et al.*, 2024]. Traditional multi-label feature selection methods primarily focus on the internal correlations within the feature set and the label set. Although some methods attempt to consider the associations between features and labels, this consideration remains insufficient [Wang *et al.*, 2021a]. For instance, linear decomposition methods such as Non-negative Matrix Factorization (NMF) may lead to the loss of nonlinear relationships within the data. This issue becomes more pronounced in real-world scenarios with complex dependencies. In the constructed composite diagram shown in Figure 1, indirect relationships between features and labels are established through the propagation of relevant nodes. Our method is specifically designed to focus on capturing these indirect relationships. To more comprehensively capture the nonlinear and indirect relationships between subsets of features and subsets of labels, we propose a novel multi-label feature selection method. The main contributions of this work are as follows:

- We propose a random walk strategy on a feature-label composite graph constructed using mutual information, which captures both direct and implicit indirect associations between features and labels, thereby facilitating a comprehensive representation of the high-dimensional structures within the data.

- By aligning the low-rank representations of similar samples in the feature space and label space, we ensure that the shared latent space after low-rank decomposition effectively captures the structural consistency between the two spaces.

- Extensive experiments conducted on seven datasets validate the effectiveness of the proposed method.

2 Related Work

In the field of multi-label feature selection, numerous methods have been proposed to capture the correlations between features and labels [Sun *et al.*, 2021]. With the increasing dimensionality of data, the diversity of labels also rises. Based on the depth of consideration given to label correlations, these methods are typically categorized into three types [Liu *et al.*, 2023], first-order strategies, second-order strategies, and high-order strategies.

First-order strategies address multi-label problems using single-label methods but ignore the correlations between labels [Lee and Kim, 2015; Lin *et al.*, 2015]. Although this method simplifies the modeling process, it fails to capture label dependencies, making it challenging to identify key features within label relationships. Second-order strategies ad-

dress this limitation to some extent by modeling pairwise relationships between labels [Xiong *et al.*, 2021; Qian *et al.*, 2022]. However, since label relationships often extend beyond simple pairwise connections, these methods only capture the associations between features and a subset of labels, falling short of comprehensively reflecting the complex dependencies between features and all labels.

In contrast, high-order strategies [Wu *et al.*, 2020; Huang *et al.*, 2016] go further by considering the correlations between features and all possible label subsets, thus providing a more comprehensive modeling of the complex dependencies among labels. However, these methods focus on uncovering the direct dependencies between variables and fail to effectively capture implicit indirect correlations, which limits their ability to characterize deep, latent relationships.

Manifold learning [Fan *et al.*, 2021; Cohen *et al.*, 2023] has been employed to ensure the consistency between feature correlation and label correlation. This method assumes that samples with similar features should have similar labels, thereby achieving alignment between feature and label similarity. The core idea is to uncover the intrinsic structure of high-dimensional data through its low-dimensional latent representations. [Tang *et al.*, 2019] proposed a graph-based manifold regularization method to capture the manifold structure of the data, applying the resulting latent representations to unsupervised feature selection. However, these methods fail to adequately address the alignment problem between the low-dimensional representations of the feature space and the label space.

In recent years, random walk techniques have seen increasing exploration and application in the field of feature selection. The Random Walk Feature Selection (RWFS) method [Feng *et al.*, 2017] is based on the Optimal Feature Selection (OPFS) technique, which identifies feature subsets and integrates random walk methods with predefined thresholds to filter out redundant features. The Pattern-Based Local Random Walk (PBLRW) method [Song *et al.*, 2019] constructs local random walk models through feature combinations and assigns transition probabilities based on the strength of these combinations. However, the random walk operations in these methods remain limited to the feature space, neglecting label dependencies, and fail to fully leverage their capacity for capturing higher-order relationships.

3 The Proposed Method

We propose a multi-label feature selection method based on Graph Random Walk and Structured Correlation Matrix Factorization (GRW-SCMF). By incorporating graph random walk, the method captures both the direct correlations and implicit indirect correlations between features and labels. Furthermore, in the shared low-dimensional representation, it aligns similar samples in the feature space and label space to ensure consistency. The framework is shown in Figure 2.

3.1 Random Walk Graph

In this paper, we construct a composite graph $G = (V, E)$, where the vertex set V consists of two types of vertices: feature vertices V_f and label vertices V_l. The edge set E includes the following three types of edges:

1. Feature-feature edges E_{ff},
2. Label-label edges E_{ll},
3. Feature-label edges E_{fl}.

The adjacency matrices of the feature graph A_{features} and the label graph A_{labels} are calculated using the Gaussian kernel as follows:

$$A_{\text{features}}(i,j) = \exp\left(-\frac{\|X_{:,i} - X_{:,j}\|^2}{2\sigma^2}\right), \qquad (1)$$

$$A_{\text{labels}}(i,j) = \exp\left(-\frac{\|Y_{:,i} - Y_{:,j}\|^2}{2\sigma^2}\right). \qquad (2)$$

Here, $\|X_{:,i} - X_{:,j}\|^2$ and $\|Y_{:,i} - Y_{:,j}\|^2$ represent the Euclidean distances between features and labels, respectively, and σ is the scale parameter of the Gaussian kernel.

The association between features and labels is represented by the mutual information matrix MI, which is defined as follows:

$$MI(i,j) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right), \qquad (3)$$

where $p(x,y)$ denotes the joint probability distribution of feature X and label Y, and $p(x)$ and $p(y)$ represent the marginal probability distributions of X and Y, respectively.

To construct the transition probability matrices P_{features}, P_{labels}, and P_{lf} (the transpose of which is P_{fl}), we normalize the adjacency matrices of the feature graph A_{features} and label graph A_{labels}, as well as the connection matrix MI between features and labels.

3.2 Random Walk Method

The random walk starts from a randomly selected feature vertex $v \in V_f$ and updates its state according to the following rules:

1. **When the current node is a feature node ($v \in V_f$):**
- With a probability of p_{jump}, it jumps to a label node $u \in V_l$. The target node is determined based on the transition probability $P_{fl}(v,u)$, which is calculated as follows:
$$P(v \to u) = p_{\text{jump}} \cdot P_{fl}(v,u), \quad u \in V_l. \qquad (4)$$

- Otherwise, with a probability of $1 - p_{\text{jump}}$, it jumps to another feature node $u \in V_f$. The target node is determined based on the transition probability $P_{\text{features}}(v,u)$, which is calculated as follows:
$$P(v \to u) = (1 - p_{\text{jump}}) \cdot P_{\text{features}}(v,u), \quad u \in V_f. \qquad (5)$$

2. **When the current node is a label node ($v \in V_l$):**
- With a probability of p_{jump}, it jumps to a feature node $u \in V_f$. The target node is determined based on the transition probability $P_{lf}(v,u)$, which is calculated as follows:
$$P(v \to u) = p_{\text{jump}} \cdot P_{lf}(v,u), \quad u \in V_f. \qquad (6)$$

- Otherwise, with a probability of $1 - p_{\text{jump}}$, it jumps to another label node $u \in V_l$. The target node is determined based on the transition probability $P_{\text{labels}}(v,u)$, which is calculated as follows:
$$P(v \to u) = (1 - p_{\text{jump}}) \cdot P_{\text{labels}}(v,u), \quad u \in V_l. \qquad (7)$$
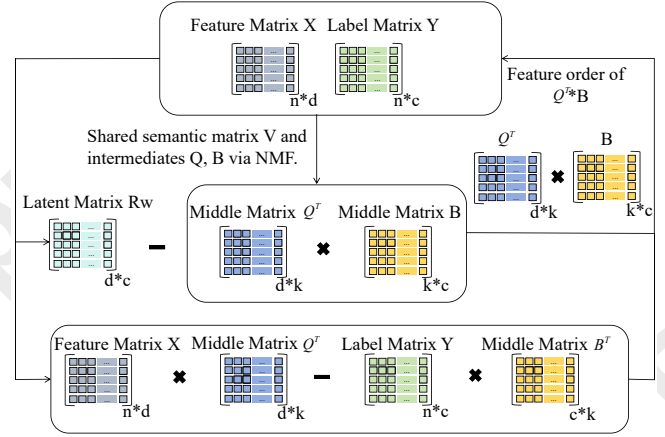

Figure 2: Algorithm framework. First, X and Y are decomposed using low-rank matrix factorization. Next, the feature-label association matrix is constrained by the random walk matrix R_w. Then, X and Y are aligned in a shared space. Finally, feature importance is computed via $Q^\top B$.

3.3 Feature-Label Relationship Update Rule

Implicit indirect relationships refer to the potential high-order associations between feature nodes and label nodes that are transmitted through intermediate nodes (other features and labels). These relationships are difficult to capture using traditional linear decomposition methods. To address this issue, RWMI leverages random walks to dynamically generate interaction sequences between feature and label nodes, effectively capturing such indirect relationships.

Specifically, during each random walk, we record each pair of feature nodes $f \in V_f$ and label nodes $l \in V_l$, and calculate their distance $d(f,l)$ in the walk sequence, which represents the number of steps separating them. The association weight $RW(f,l)$ is updated according to the following formula:

$$RW(f,l) + = \text{decay\_factor}^{d(f,l)} \cdot MI(f,l), \qquad (8)$$

where decay_factor is a distance decay factor that controls the influence of the number of steps on the weight, and $MI(f,l)$ is the mutual information between feature f and label l, which mitigates uncertainties caused by randomness, ensuring that the results of the RW matrix are more accurate and robust.

Through multiple random walks, RW not only captures the direct associations between features and labels but also computes implicit indirect relationships via the intermediate nodes in the walk sequences. This process constructs a feature-label association matrix that represents high-order associations, which is further scaled to the $[0,1]$ range via Min–Max normalization.

The procedure is detailed in Algorithm 1.

3.4 Objective Function

In multi-label data, the diversity of labels may introduce inaccurate or noisy label information, making it prone to incorporating irrelevant and redundant information when directly utilizing the original feature and label matrices for feature selec-

---

Algorithm 1 Random Walk Mutual Information (RWMI)

Require: X (features), Y (labels), n_walks, $walk\_length$, $jump\_prob$, $decay\_factor$
Ensure: Normalized RWMI matrix
 1: Construct adjacency matrices $A_{features}$, A_{labels} from X, Y.
 2: Calculate and normalize mutual information matrix MI between X and Y.
 3: Compute transition probability matrices: $P_{features}$, P_{labels}, $P_{feature\_to\_label}$, $P_{label\_to\_feature}$.
 4: Initialize $RWMI$ matrix.
 5: **for** $i = 1$ to n_walks **do**
 6:     Start random walk from a node.
 7:     **for** $j = 1$ to $walk\_length$ **do**
 8:         Move between feature and label nodes based on probabilities and $jump\_prob$.
 9:     **end for**
10:     Update $RWMI$ matrix using decay factor and steps between feature-label pairs, and normalize it afterward.
11: **end for**

---

tion. To address this issue, we learn a shared low-dimensional latent semantic space for both the feature and label spaces. This low-dimensional space matrix condenses the critical information from features and labels while effectively reducing the influence of redundancy and noise. The objective function can be written as:

$$\min_{V,Q,B} \|X - VQ\|_F^2 + \|Y - VB\|_F^2 \tag{9}$$

The feature matrix $X \in \mathbb{R}^{n \times d}$ represents the high-dimensional feature space of the samples, while the multi-label matrix $Y \in \mathbb{R}^{n \times c}$ describes the label distribution of the samples. To uncover the shared distribution structure between the feature space and the label space, we introduce a low-dimensional latent semantic matrix $V \in \mathbb{R}^{n \times k}$, which captures the low-dimensional shared representations of both spaces. The matrices $Q \in \mathbb{R}^{k \times d}$ and $B \in \mathbb{R}^{k \times c}$ are the low-dimensional representation coefficient matrices for features and labels, respectively, mapping the original data into the shared representation V. Specifically, matrix V encodes the embedding relationships of features and labels in k latent semantic clusters. By applying non-negative matrix factorization (NMF) on Y, the c labels can be effectively clustered, revealing the complex internal distribution structure of multi-label data.

$$\min_{V,Q,B} \|X - VQ\|_F^2 + \|Y - VB\|_F^2 + \|R_w - Q^T B\|_F^2 \tag{10}$$

The term $\|R_w - Q^T B\|_F^2$ introduced in the objective function (10) enhances the model's ability to capture the complex relationships between features and labels. Here, the matrix R_w is the feature-label implicit association matrix computed using the Random Walk Mutual Information (RWMI) method, which effectively captures indirect relations and latent dependencies in high-dimensional data. In contrast, $Q^T B$ is the feature-label association matrix learned based on the low-dimensional shared semantic space. As it is derived from low-dimensional linear decomposition, it primarily captures linear relationships and direct associations. By combining the explicit modeling capability of linear decomposition with the implicit relationship mining power of random walks, this term significantly improves the model's ability to represent high-order nonlinear relationships between features and labels. Furthermore, this term ensures that the low-dimensional representation space preserves the manifold structure of the original high-dimensional data, aligning with the latest concepts in manifold learning, which posit that high-dimensional data often lie on a low-dimensional manifold.

$$\min_{V,Q,B} \|X - VQ\|_F^2 + \|Y - VB\|_F^2 + \|R_w - Q^T B\|_F^2 + \|XQ^T - YB^T\|_F^2 \tag{11}$$

This term can be regarded as an alignment constraint, aiming to ensure the consistency between the feature space and the label space in the low-dimensional representation. Specifically, XQ^T represents the low-dimensional mapping of features in the shared semantic space V, while YB^T represents the low-dimensional mapping of labels in the same space V. Furthermore, to preserve the associations between features and labels, similar feature samples and label samples should be mapped to locations close to each other in the low-dimensional space. This ensures the learning of a common low-dimensional space that effectively captures the relationship between features and labels.

$$\min_{V,Q,B} \|X - VQ\|_F^2 + \|Y - VB\|_F^2 + \|R_w - Q^T B\|_F^2 + \|XQ^T - YB^T\|_F^2 + \|Q^T B\|_{2,1} + \|V\|_F^2 \tag{12}$$

The Frobenius norm of V imposes a regularization constraint in the low-dimensional shared semantic space, reducing redundancy and unnecessary complexity, thereby enhancing the model's ability to capture important low-dimensional relationships between features and labels. Moreover, incorporating this term as a regularization on $Q^T B$ effectively controls the model complexity and reduces the risk of overfitting. By smoothing the solution space, it further improves the generalization ability of the model. Additionally, computing the row-wise ℓ_2 norm of $Q^T B$ provides a measure of feature importance.

Thus, our final objective function is formulated as follows:

$$\min_{V,Q,B} \alpha\|X - VQ\|_F^2 + \beta\|Y - VB\|_F^2 + \gamma\|R_w - Q^T B\|_F^2 + \delta\|XQ^T - YB^T\|_F^2 + \epsilon\|Q^T B\|_{2,1} + \|V\|_F^2 \tag{13}$$
$$\text{s.t. } \{V, Q, B\} \geq 0$$

4 Solution Strategy

This section presents the optimization method for solving the proposed objective function.

4.1 Optimization Scheme

Because the objective function includes an $l_{2,1}$-norm regularization term, which is non-smooth and cannot be directly solved, and is also non-convex with respect to the variables V, Q, and B, it presents a significant challenge [Boyd and Vandenberghe, 2004]. Theoretically, this is due to the fact that the

Hessian matrix, composed of the second-order partial derivatives of the objective function, is not positive semi-definite. Therefore, to optimize the objective function, we have designed a relaxation update method based on the alternating-multiplier-based scheme [Nie *et al.*, 2010] to achieve an approximate local optimum of the objective function.

$$\min_{V,Q,B} \alpha\|X - VQ\|_F^2 + \beta\|Y - VB\|_F^2 + \gamma\|R_w - Q^T B\|_F^2$$
$$+ \delta\|XQ^T - YB^T\|_F^2 + \|V\|_F^2 + 2\epsilon\,\mathrm{Tr}(W^T DW) \quad (14)$$

We use $2\,\mathrm{Tr}(W^T DW)$ to approximate $\|W\|_{2,1}$, where $W = Q^T B$. Here, D is a diagonal matrix whose elements are iteratively computed during the method's execution. The parameter c is introduced to mitigate perturbations in non-differentiable problems. The elements of D are defined as:

$$D_{ii} = \frac{1}{2\sqrt{W_i^T W_i + c}}, \quad (c \to 0) \quad (15)$$

To integrate non-negative constraint conditions into the objective function, we introduce Lagrangian multipliers ψ, φ, and μ to constrain V, Q, and B respectively. Specifically, $\psi \in \mathbb{R}_+^{n \times k}$, $\varphi \in \mathbb{R}_+^{k \times d}$, and $\mu \in \mathbb{R}_+^{k \times c}$. Consequently, the original function (14) is equivalent to the following function:

$$\min_{V,Q,B} \alpha\|X - VQ\|_F^2 + \beta\|Y - VB\|_F^2 + \gamma\|R_w - Q^T B\|_F^2$$
$$+ \delta\|XQ^T - YB^T\|_F^2 + \|V\|_F^2 + 2\epsilon\,\mathrm{Tr}(W^T DW) \quad (16)$$
$$- \mathrm{Tr}(\psi V^T) - \mathrm{Tr}(\varphi Q^T) - \mathrm{Tr}(\mu B^T)$$

Transform the squared Frobenius norm into the form of a matrix trace:

$$\begin{aligned}
\Theta = \ &\alpha\,\mathrm{Tr}\left((X - VQ)^T(X - VQ)\right) \\
&+ \beta\,\mathrm{Tr}\left((Y - VB)^T(Y - VB)\right) \\
&+ \gamma\,\mathrm{Tr}\left((R_w - Q^T B)^T(R_w - Q^T B)\right) \\
&+ \delta\,\mathrm{Tr}\left((XQ^T - YB^T)^T(XQ^T - YB^T)\right) \\
&+ \mathrm{Tr}(V^T V) + 2\epsilon\,\mathrm{Tr}(W^T DW) \\
&- \mathrm{Tr}(\psi V^T) - \mathrm{Tr}(\varphi Q^T) - \mathrm{Tr}(\mu B^T).
\end{aligned} \quad (17)$$

We can obtain the following expressions by differentiating with respect to V, Q, and B:

$$\begin{cases}
\dfrac{\partial\Theta}{\partial V} = 2(\alpha VQQ^T + \beta VBB^T + V - \alpha XQ^T - \beta YB^T) - \psi \\[2mm]
\dfrac{\partial\Theta}{\partial Q} = 2(\alpha V^T VQ + \gamma BB^T Q + \epsilon BB^T QD + \delta QX^T X \\
\qquad\qquad - \alpha V^T X - \gamma BR_w^T - \delta BY^T X) - \varphi \\[2mm]
\dfrac{\partial\Theta}{\partial B} = 2(\beta V^T VB + \gamma QQ^T B + \epsilon QDQ^T B + \delta BY^T Y \\
\qquad\qquad - \beta V^T Y - \gamma QR_w - \delta QX^T Y) - \mu
\end{cases} \quad (18)$$

By employing the Karush-Kuhn-Tucker (KKT) conditions for optimization, we derive the following set of equations:

$$\begin{cases}
(\alpha VQQ^T + \beta VBB^T + V - \alpha XQ^T - \beta YB^T) \circ V = 0 \\[2mm]
(\alpha V^T VQ + \gamma BB^T Q + \epsilon BB^T QD + \delta QX^T X \\
\quad - \alpha V^T X - \gamma BR_w^T - \delta BY^T X) \circ Q = 0 \\[2mm]
(\beta V^T VB + \gamma QQ^T B + \epsilon QDQ^T B + \delta BY^T Y \\
\quad - \beta V^T Y - \gamma QR_w - \delta QX^T Y) \circ B = 0
\end{cases} \quad (19)$$

Dataset	#Training set	#Test set	#Features	#Labels	#Distinct	#Domain
Arts	2000	3000	462	26	321 ± 139	Web text
Business	2000	3000	438	30	321 ± 139	Web text
Education	2000	3000	550	33	321 ± 139	Web text
Health	2000	3000	612	32	321 ± 139	Web text
Yeast	1500	917	103	14	198	Biology
Flags	129	65	19	7	54	Image
Emotions	391	202	72	6	27	Music

Table 1: Elaborated information regarding the experimental datasets.

In the case of multiple variables with constraints on each variable, we alternately update each variable to ensure that each optimization step adheres to the constraint conditions. We take the partial derivatives with respect to V, Q, and B respectively.

$$V \leftarrow V \odot \frac{\alpha XQ^T + \beta YB^T}{\alpha VQQ^T + \beta VBB^T + V} \quad (20)$$

$$Q \leftarrow Q \odot \frac{\alpha V^T X + \gamma BR_w^T + \delta BY^T X}{\alpha V^T VQ + \gamma BB^T Q + \epsilon BB^T QD + \delta QX^T X} \quad (21)$$

$$B \leftarrow B \odot \frac{\beta V^T Y + \gamma QR_w + \delta QX^T Y}{\beta V^T VB + \gamma QQ^T B + \epsilon QDQ^T B + \delta BY^T Y} \quad (22)$$

After obtaining each parameter of the objective function, we can rank all features of X in descending order based on the values of $\|(Q^\top B)_i\|_2$ (for $i = 1, \ldots, d$).

5 Experiments

5.1 Experimental Setup

Datasets. To validate the effectiveness of the proposed method under complex relationships, our experimental evaluation employs multi-label datasets from the MULAN library [Tsoumakas *et al.*, 2011], including the web text datasets Arts, Business, Education, and Health. Additionally, the datasets encompass the Emotions dataset for music genres, the Flags dataset for image data, and the Yeast dataset for biological data. The Arts, Business, Education, and Health datasets feature a large number of distinct label combinations, reflecting complex underlying relationships with the features. Meanwhile, the Emotions, Flags, and Yeast datasets highlight characteristics from different domains. Table 1 summarizes the characteristics of the datasets used in the experiments.

Comparing Methods. We compare a broad and representative set of multi-label feature selection methods: (1) **Statistical methods**: PPT+MI [Doquire and Verleysen, 2011] and PPT+CHI [Read, 2008], which perform feature selection based on mutual information and χ^2 statistics, respectively; (2) **Latent representation learning**: LRDG [Zhang *et al.*, 2024], which incorporates dynamic graph constraints; (3) **Label relevance-based methods**: MIFS [Jian *et al.*, 2016], which selects features in a low-dimensional space, and LRFS [Zhang *et al.*, 2019], which leverages conditional mutual information to capture label relationships; (4) **Sparse constraint-based methods**: RALM-FS [Cai *et al.*, 2013], which employs the $l_{2,0}$-norm for sparse solutions; (5) **Latent shared structure-based methods**: SSFS [Gao *et al.*, 2023], which models the shared latent structure between features and labels.

Dataset	GRW-SCMF	PPT+MI	PPT+CHI	MIFS	LRFS	RALM-FS	SSFS	LRDG
				Micro-F_1				
Flags	**0.7403±0.010**	0.6646±0.040	0.6632±0.038	0.7225±0.050	0.6659±0.041	0.6315±0.048	0.6520±0.034	0.7040±0.047
Emotions	**0.5833±0.066**	0.3442±0.175	0.2730±0.138	0.0684±0.059	0.2257±0.185	0.0082±0.030	0.4215±0.117	0.3223±0.147
Yeast	**0.5900±0.028**	0.5568±0.028	0.5622±0.032	0.5659±0.028	0.5523±0.030	0.5590±0.029	0.5395±0.033	0.5677±0.038
Arts	**0.2823±0.061**	0.0904±0.053	0.0981±0.055	0.1391±0.078	0.1041±0.045	0.1018±0.061	0.1878±0.092	0.0584±0.024
Business	**0.6986±0.011**	0.6729±0.004	0.6725±0.004	0.6835±0.008	0.6796±0.005	0.6696±0.001	0.6817±0.010	0.6723±0.005
Education	**0.3135±0.060**	0.1237±0.083	0.1199±0.085	0.0733±0.059	0.1495±0.081	0.1934±0.056	0.2500±0.084	0.2283±0.086
Health	**0.5549±0.033**	0.4017±0.074	0.3993±0.075	0.4681±0.080	0.4754±0.037	0.5157±0.044	0.4999±0.102	0.4419±0.065
				Macro-F_1				
Flags	**0.5992±0.015**	0.5135±0.037	0.5174±0.046	0.5697±0.100	0.5106±0.044	0.4872±0.044	0.4987±0.048	0.5486±0.089
Emotions	**0.5213±0.068**	0.2825±0.145	0.1925±0.118	0.0409±0.037	0.1660±0.155	0.0051±0.018	0.2521±0.070	0.2798±0.142
Yeast	**0.2830±0.038**	0.2337±0.038	0.2435±0.046	0.2511±0.039	0.2280±0.045	0.2397±0.042	0.2084±0.044	0.2504±0.050
Arts	**0.1162±0.030**	0.0373±0.022	0.0397±0.023	0.0550±0.034	0.0433±0.019	0.0385±0.026	0.0790±0.038	0.0222±0.009
Business	**0.1086±0.026**	0.0385±0.008	0.0395±0.008	0.0564±0.011	0.0547±0.007	0.0320±0.001	0.0546±0.014	0.0424±0.010
Education	**0.0855±0.019**	0.0342±0.025	0.0345±0.026	0.0195±0.017	0.0485±0.028	0.0525±0.014	0.0647±0.020	0.0632±0.027
Health	**0.1670±0.042**	0.1023±0.044	0.1039±0.043	0.1181±0.046	0.1454±0.038	0.1592±0.042	0.1658±0.063	0.1038±0.055

Table 2: The classification performance of all methodologies, evaluated using the SVM classifier, is provided in terms of Micro-F_1 and Macro-F_1 (mean±std).

Evaluating Methods. Analogous to method [Hu *et al.*, 2020], we select the top 20% of ranked features from each dataset, with a step size of 1% (for the Flags dataset, all features were used as it only has 19 features). Three commonly used classifiers were employed: Linear SVM, k-Nearest Neighbors (k=3), and MLkNN (k=10). Specifically, both the Linear SVM and kNN classifiers provided Micro-F_1 and Macro-F_1 scores, while the MLkNN classifier yields Hamming Loss (HL) and Zero-One Loss (ZOL) [Kou *et al.*, 2023; Li *et al.*, 2023].

Parameter Selection. In this method, we employ Bayesian optimization to determine the optimal combination of parameters. The search range for the regularization parameters is set to {0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0}. For the random walk method, the range for n_{walks} is set to {100, 1000, 10000}, the range for $walk\_length$ is set to {10, 20, 30}, and the ranges for $jump\_prob$ and $decay\_factor$ are both set to {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}.

5.2 Experimental Results

Feature Selection Performance. We selected seven benchmark multi-label datasets for our experiments and compared our method with seven state-of-the-art multi-label feature selection methods. The experimental results are presented in the tables, where our method outperforms the others. In Tables 2 and 3, the best-performing method for each dataset is highlighted in bold. Clearly, our method is not limited to any specific type of dataset and demonstrates outstanding performance across image, music, text, and biological datasets. Except for the Macro-F_1 score of the 3NN classifier on the Health dataset, where our method slightly underperforms compared to the RALM-FS method, our method outperforms the seven other methods in all other scenarios. In Figure 3, we present the performance of different methods on a specific dataset. From the early stages, the proposed method demonstrates its ability to select highly discriminative fea-
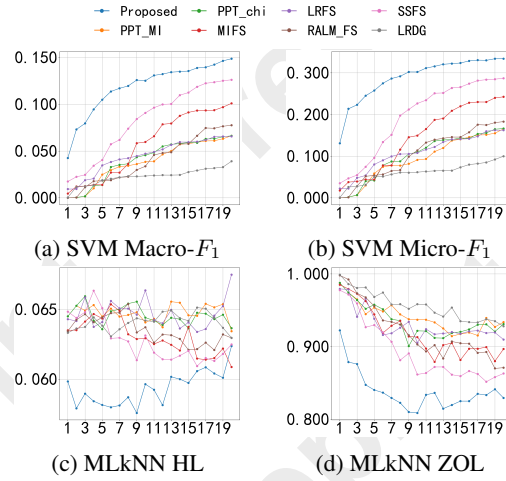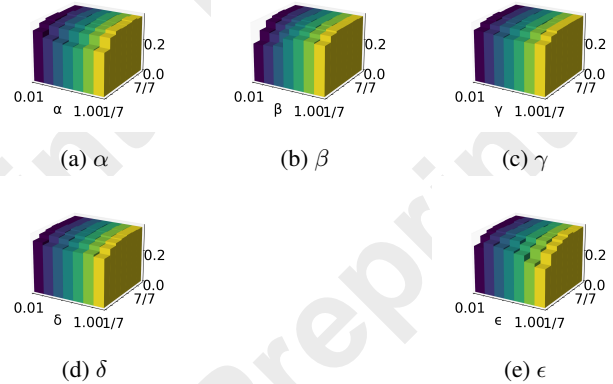
(a) SVM Macro-F_1 (b) SVM Micro-F_1

(c) MLkNN HL (d) MLkNN ZOL

Figure 3: Eight methods on Arts.

(a) α (b) β (c) γ

(d) δ (e) ϵ

Figure 4: Parameter sensitivity studies on the Arts dataset.

Dataset	GRW-SCMF	PPT+MI	PPT+CHI	MIFS	LRFS	RALM-FS	SSFS	LRDG
				Micro-F_1				
Flags	**0.6871±0.012**	0.6150±0.011	0.6154±0.011	0.6615±0.030	0.6157±0.018	0.6057±0.012	0.6133±0.014	0.6632±0.037
Emotions	**0.5977±0.061**	0.5324±0.038	0.5276±0.049	0.4877±0.074	0.4771±0.032	0.5492±0.075	0.3701±0.048	0.4452±0.069
Yeast	**0.5725±0.027**	0.5494±0.025	0.5477±0.023	0.5547±0.021	0.5406±0.027	0.5477±0.024	0.3131±0.132	0.5553±0.032
Arts	**0.2856±0.023**	0.1812±0.037	0.1870±0.034	0.2016±0.052	0.1963±0.027	0.1824±0.044	0.2300±0.049	0.1272±0.040
Business	**0.6745±0.011**	0.6515±0.023	0.6556±0.023	0.6605±0.066	0.6700±0.009	0.6300±0.077	0.6626±0.020	0.6598±0.015
Education	**0.3122±0.027**	0.2301±0.043	0.2274±0.043	0.1827±0.055	0.2431±0.046	0.2605±0.051	0.2856±0.043	0.2769±0.054
Health	**0.4878±0.012**	0.3905±0.044	0.3870±0.045	0.4350±0.071	0.4365±0.020	0.4739±0.059	0.4537±0.052	0.4256±0.080
				Macro-F_1				
Flags	**0.5782±0.011**	0.4480±0.011	0.4493±0.011	0.5265±0.078	0.4531±0.022	0.4386±0.011	0.4504±0.020	0.5179±0.069
Emotions	**0.5824±0.064**	0.5143±0.040	0.4997±0.051	0.4746±0.075	0.4551±0.030	0.5355±0.076	0.1751±0.028	0.4274±0.072
Yeast	**0.3718±0.030**	0.3404±0.027	0.3376±0.025	0.3407±0.022	0.3287±0.028	0.3421±0.025	0.1380±0.049	0.3287±0.059
Arts	**0.1261±0.027**	0.0879±0.022	0.0912±0.022	0.0951±0.034	0.1006±0.025	0.0714±0.025	0.1186±0.035	0.0534±0.016
Business	**0.1146±0.022**	0.0810±0.019	0.0785±0.015	0.1075±0.023	0.0958±0.017	0.0654±0.018	0.0985±0.025	0.0931±0.026
Education	**0.1002±0.015**	0.0816±0.019	0.0829±0.022	0.0426±0.018	0.0870±0.020	0.0838±0.023	0.0959±0.020	0.0865±0.029
Health	0.1703±0.024	0.1329±0.029	0.1327±0.027	0.1570±0.041	0.1611±0.023	**0.1859±0.032**	0.1738±0.037	0.1302±0.053

Table 3: The classification performance of all methodologies, assessed using the 3NN classifier, is presented in terms of Micro-F_1 and Macro-F_1 (mean±std).

RW	FLA	Emotions	Yeast	Arts
	✓	0.5532	0.5808	0.2820
✓		0.5027	0.5846	0.2084
✓	✓	0.5833	0.5900	0.2823

Table 4: Ablation experimental results.

(a) Arts (b) Business

Figure 5: Convergence curves on Arts and Business datasets.

tures. The outstanding results in HL and ZOL metrics validate its effectiveness in capturing both direct and indirect relationships between features and labels. Furthermore, the superior performance in Macro-F_1 and Micro-F_1 scores clearly highlights the efficacy of low-dimensional alignment.

Parameter Analysis. In our proposed method, there exist five weight parameters, α, β, γ, δ and ϵ, that influence performance outcomes. The Figure 4 illustrates how these five parameters affect the Micro-F_1 scores computed using SVM on the Arts dataset. We initially divided the dataset into seven equal parts and then individually adjusted each parameter while keeping the other parameters fixed at 0.5 [Jian *et al.*, 2018], conducting a grid search over a predefined range. It is evident that the results exhibit an upward trend and become increasingly stable as the number of selected features increases. However, due to different parameter configurations, slight variations may occur under the same number of selected features, leading to minor discrepancies in the results. Therefore, we can infer that our method demonstrates stability, provided there is a sufficiently large number of instances.

Ablation Study. To evaluate the effectiveness of the newly introduced components, we conducted ablation experiments on the Emotions, Yeast, and Arts datasets, considering various combinations of these components, as shown in Table 4. In these experiments, the random walk component (RW) captures implicit indir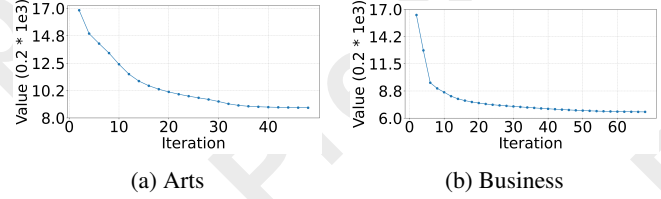ect relationships through random walk, while the feature-label alignment component (FLA) aligns features and labels in a low-dimensional common space. The results demonstrate that removing either component adversely affects the SVM's Micro-F_1 performance, underscoring the importance of both introduced components. Therefore, it can be inferred that the contribution of each component is crucial for achieving optimal model performance.

Convergence. Figure 5 illustrates the convergence of our method on the Arts and Business datasets. The chosen stopping criterion is: $|z_t - z_{t-1}| < c$ or $\left|\frac{z_t - z_{t-1}}{z_{t-1}}\right| < c$. The X-axis represents the number of iterations, and the Y-axis is similar to the stopping criterion but uses absolute values. It is evident from the figure that the optimization process converges quickly.

6 Conclusion

We propose a random walk method based on a feature-label composite graph and incorporate it into a multi-label feature selection method. Through random walks on the feature-label composite graph, we capture direct and indirect correlation between features and labels. Additionally, we leverage low-dimensional representation coefficients to align the low-dimensional variable space while preserving the manifold structure. Experimental results demonstrate the effectiveness and robustness of our method.

Acknowledgments

References

[Al-Yaseen *et al.*, 2022] Wathiq Laftah Al-Yaseen, Ali Kadhum Idrees, and Faezah Hamad Almasoudy. Wrapper feature selection method based differential evolution and extreme learning machine for intrusion detection system. *Pattern Recognition*, 132:108912, 2022.

[Bommert *et al.*, 2022] Andrea Bommert, Thomas Welchowski, Matthias Schmid, and Jörg Rahnenführer. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings in Bioinformatics*, 23(1):bbab354, 2022.

[Boyd and Vandenberghe, 2004] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Cai *et al.*, 2013] Xiao Cai, Feiping Nie, and Heng Huang. Exact top-k feature selection via l2, 0-norm constraint. In *Twenty-third international joint conference on artificial intelligence*. Citeseer, 2013.

[Cohen *et al.*, 2023] David Cohen, Tal Shnitzer, Yuval Kluger, and Ronen Talmon. Few-sample feature selection via feature manifold learning. In *International Conference on Machine Learning*, pages 6296–6319. PMLR, 2023.

[Doquire and Verleysen, 2011] Gauthier Doquire and Michel Verleysen. Feature selection for multi-label classification problems. In *Advances in Computational Intelligence: 11th International Work-Conference on Artificial Neural Networks, IWANN 2011, Torremolinos-Málaga, Spain, June 8-10, 2011, Proceedings, Part I 11*, pages 9–16. Springer, 2011.

[Fan *et al.*, 2021] Yuling Fan, Jinghua Liu, Peizhong Liu, Yongzhao Du, Weiyao Lan, and Shunxiang Wu. Manifold learning with structured subspace for multi-label feature selection. *Pattern Recognition*, 120:108169, 2021.

[Feng *et al.*, 2017] Lizhou Feng, Youwei Wang, and Wanli Zuo. Novel feature selection method based on random walk and artificial bee colony. *Journal of Intelligent & Fuzzy Systems*, 32(1):115–126, 2017.

[Gao *et al.*, 2023] Wanfu Gao, Yonghao Li, and Liang Hu. Multilabel feature selection with constrained latent structure shared term. *IEEE Transactions on Neural Networks and Learning Systems*, 34(3):1253–1262, 2023.

[Hu *et al.*, 2020] Liang Hu, Yonghao Li, Wanfu Gao, Ping Zhang, and Juncheng Hu. Multi-label feature selection with shared common mode. *Pattern Recognition*, 104:107344, 2020.

[Huang *et al.*, 2016] Jun Huang, Guorong Li, Qingming Huang, and Xindong Wu. Learning label-specific features and class-dependent labels for multi-label classification.

IEEE transactions on knowledge and data engineering, 28(12):3309–3323, 2016.

[Jian *et al.*, 2016] Ling Jian, Jundong Li, Kai Shu, and Huan Liu. Multi-label informed feature selection. In *IJCAI*, volume 16, pages 1627–33, 2016.

[Jian *et al.*, 2018] Ling Jian, Jundong Li, and Huan Liu. Exploiting multilabel information for noise-resilient feature selection. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–23, 2018.

[Jin *et al.*, 2023] Lingbin Jin, Li Zhang, and Lei Zhao. Feature selection based on absolute deviation factor for text classification. *Information Processing & Management*, 60(3):103251, 2023.

[Kong *et al.*, 2012] Deguang Kong, Chris Ding, Heng Huang, and Haifeng Zhao. Multi-label relieff and f-statistic feature selections for image annotation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2352–2359. IEEE, 2012.

[Kou *et al.*, 2023] Yi Kou, Guoping Lin, Yuhua Qian, and Shujiao Liao. A novel multi-label feature selection method with association rules and rough set. *Information Sciences*, 624:299–323, 2023.

[Lee and Kim, 2015] Jaesung Lee and Dae-Won Kim. Mutual information-based multi-label feature selection using interaction information. *Expert Systems with Applications*, 42(4):2013–2025, 2015.

[Lee and Kim, 2017] Jaesung Lee and Dae-Won Kim. Scls: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognition*, 66:342–352, 2017.

[Li *et al.*, 2017] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[Li *et al.*, 2023] Yonghao Li, Liang Hu, and Wanfu Gao. Multi-label feature selection via robust flexible sparse regularization. *Pattern Recognition*, 134:109074, 2023.

[Li *et al.*, 2024] Yonghao Li, Liang Hu, and Wanfu Gao. Multi-label feature selection with high-sparse personalized and low-redundancy shared common features. *Information Processing & Management*, 61(3):103633, 2024.

[Lim and Kim, 2020] Hyunki Lim and Dae-Won Kim. Mfc: Initialization method for multi-label feature selection based on conditional mutual information. *Neurocomputing*, 382:40–51, 2020.

[Lin *et al.*, 2015] Yaojin Lin, Qinghua Hu, Jinghua Liu, and Jie Duan. Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing*, 168:92–103, 2015.

[Liu *et al.*, 2023] Bo Liu, Weibin Li, Yanshan Xiao, Xiaodong Chen, Laiwang Liu, Changdong Liu, Kai Wang, and Peng Sun. Multi-view multi-label learning with high-order label correlation. *Information Sciences*, 624:165–184, 2023.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint ℓ_2, 1-norms minimization. *Advances in neural information processing systems*, 23, 2010.

[Qian *et al.*, 2022] Wenbin Qian, Chuanzhen Xiong, Yuhua Qian, and Yinglong Wang. Label enhancement-based feature selection via fuzzy neighborhood discrimination index. *Knowledge-Based Systems*, 250:109119, 2022.

[Read, 2008] Jesse Read. A pruned problem transformation method for multi-label classification. In *Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008)*, volume 143150, page 41, 2008.

[Silla Jr *et al.*, 2008] Carlos N Silla Jr, Alessandro L Koerich, and Celso AA Kaestner. Feature selection in automatic music genre classification. In *2008 Tenth IEEE international symposium on multimedia*, pages 39–44. IEEE, 2008.

[Song *et al.*, 2019] Aibo Song, Yangyang Liu, Zhiang Wu, Mingyu Zhai, and Junzhou Luo. A local random walk model for complex networks based on discriminative feature combinations. *Expert Systems with Applications*, 118:329–339, 2019.

[Sun *et al.*, 2021] Lin Sun, Tianxiang Wang, Weiping Ding, Jiucheng Xu, and Yaojin Lin. Feature selection using fisher score and multilabel neighborhood rough sets for multilabel classification. *Information Sciences*, 578:887–912, 2021.

[Tang *et al.*, 2019] Chang Tang, Meiru Bian, Xinwang Liu, Miaomiao Li, Hua Zhou, Pichao Wang, and Hailin Yin. Unsupervised feature selection via latent representation learning and manifold regularization. *Neural Networks*, 117:163–178, 2019.

[Tang *et al.*, 2021] Jing Tang, Minjie Mou, Yunxia Wang, Yongchao Luo, and Feng Zhu. Metafs: performance assessment of biomarker discovery in metaproteomics. *Briefings in Bioinformatics*, 22(3):bbaa105, 2021.

[Tsoumakas *et al.*, 2011] Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, 12:2411–2414, 2011.

[Wang *et al.*, 2021a] Qianqian Wang, Huanhuan Lian, Gan Sun, Quanxue Gao, and Licheng Jiao. icmsc: Incomplete cross-modal subspace clustering. *IEEE Transactions on Image Processing*, 30:305–317, 2021.

[Wang *et al.*, 2021b] Ran Wang, Sam Kwong, Xu Wang, and Yuheng Jia. Active k-labelsets ensemble for multi-label classification. *Pattern Recognition*, 109:107583, 2021.

[Wu *et al.*, 2020] Xingyu Wu, Bingbing Jiang, Kui Yu, Huanhuan Chen, and Chunyan Miao. Multi-label causal feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6430–6437, 2020.

[Xiong *et al.*, 2021] Chuanzhen Xiong, Wenbin Qian, Yinglong Wang, and Jintao Huang. Feature selection based on label distribution and fuzzy mutual information. *Information Sciences*, 574:297–319, 2021.

[Yamada *et al.*, 2020] Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. In *International conference on machine learning*, pages 10648–10659. PMLR, 2020.

[Yao *et al.*, 2017] Chao Yao, Ya-Feng Liu, Bo Jiang, Jungong Han, and Junwei Han. Lle score: A new filter-based unsupervised feature selection method based on nonlinear manifold embedding and its application to image recognition. *IEEE Transactions on Image Processing*, 26(11):5257–5269, 2017.

[Yu *et al.*, 2021] Kui Yu, Mingzhu Cai, Xingyu Wu, Lin Liu, and Jiuyong Li. Multilabel feature selection: a local causal structure learning approach. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6):3044–3057, 2021.

[Zhang and Gao, 2021] Ping Zhang and Wanfu Gao. Feature relevance term variation for multi-label feature selection. *Applied Intelligence*, 51:5095–5110, 2021.

[Zhang *et al.*, 2019] Ping Zhang, Guixia Liu, and Wanfu Gao. Distinguishing two types of labels for multi-label feature selection. *Pattern recognition*, 95:72–82, 2019.

[Zhang *et al.*, 2021] Ping Zhang, Wanfu Gao, Juncheng Hu, and Yonghao Li. A conditional-weight joint relevance metric for feature relevancy term. *Engineering Applications of Artificial Intelligence*, 106:104481, 2021.

[Zhang *et al.*, 2024] Yao Zhang, Wei Huo, and Jun Tang. Multi-label feature selection via latent representation learning and dynamic graph constraints. *Pattern Recognition*, 151:110411, 2024.

[Zhu *et al.*, 2018] Yue Zhu, Kai Ming Ting, and Zhi-Hua Zhou. Multi-label learning with emerging new labels. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1901–1914, 2018.