EMMS: Evidential Multi-Label Multi-Dimensional Selection

Li Yang[1] , **Yanyong Huang**[1*] , **Jinyuan Chang**[1] , **Ou Zheng**[2] , **Minbo Ma**[3] and **Xiaoyi Jiang**[4]
[1]Joint Laboratory of Data Science and Business Intelligence, School of Statistics and Data Science, Southwestern University of Finance and Economics
[2]Zhiling Research
[3]Institute for Carbon Neutrality, Tsinghua University
[4]Department of Computer Science, University of Münster
1240202j8007@smail.swufe.edu.cn, {huangyy, changjinyuan}@swufe.edu.cn,
ouzheng@zhilingresearch.com, minboma@mail.tsinghua.edu.cn, xjiang@uni-muenster.de

Abstract

Multi-label data often contain high-dimensional features, outlier instances, and noisy labels, all of which can lead to the curse of dimensionality and decreased performance in downstream tasks. Although numerous data reduction methods have been developed, existing approaches face two major limitations: 1) existing methods typically select features, instances, or labels independently, without considering how noise or redundancy in one dimension may negatively influence the selection of others; 2) there are very few feature and instance co-selection methods that commonly assume label annotations are free of noise, which is seldom true in practice. To address these issues, we propose Evidential Multi-Label Multi-Dimensional Selection (EMMS), which jointly performs feature, instance, and label selection on multi-label data. EMMS introduces a dual projection mechanism with sparsity constraints that transforms high-dimensional data first into a latent space and then into the label space. Simultaneously, projection residuals are explicitly modeled to facilitate the identification of representative instances, enabling unified selection across features, instances, and labels. Moreover, EMMS employs evidence theory to fuse instance-level and label-level evidence, thereby enhancing the reliability of the learned labels and reducing the influence of noisy labels, which in turn promotes multi-dimensional selection. Extensive experiments demonstrate that EMMS consistently outperforms state-of-the-art methods.

1 Introduction

Multi-label data, in which each instance is associated with multiple labels simultaneously, is prevalent in real-world applications [Zhang and Zhou, 2006; Lin *et al.*, 2015]. For example, in image classification tasks, a single image may simultaneously be labeled as "sky," "building," and "person" [Wang *et al.*, 2016]. Multi-label data is often high-dimensional, containing redundant or irrelevant features, and may also include noisy or outlier instances. Furthermore, because obtaining accurate label annotations for every instance is expensive and often requires expert knowledge, the resulting labels are often subject to noise. These characteristics can lead to the curse of dimensionality, increase the risk of overfitting, and ultimately degrade the performance of downstream tasks [Lin *et al.*, 2015; Li *et al.*, 2022c]. Therefore, how to select discriminative features, representative instances, and informative labels from high-dimensional multi-label data that contains noisy instances and labels has become an urgent problem in many practical applications.

To address this issue, data reduction techniques for multi-label data, an important area of research in multi-label learning, have been extensively investigated from multiple perspectives [Zhang *et al.*, 2023; Del Castillo *et al.*, 2021; Pan *et al.*, 2022]. Depending on the dimension being reduced, existing approaches for multi-label data can be broadly categorized into feature selection, instance selection, and label selection. Multi-label feature selection aims to identify a subset of informative features from multi-label data to alleviate the curse of dimensionality [Zhang *et al.*, 2023; Zhang *et al.*, 2025; Wang *et al.*, 2025], while multi-label instance selection focuses on choosing representative instances to reduce sample size [Del Castillo *et al.*, 2021; Ougiaroglou *et al.*, 2023; Li *et al.*, 2024]. Additionally, label selection for multi-label data seeks to preserve the most informative labels, thus reducing label redundancy [Pan *et al.*, 2022; Nguyen *et al.*, 2024]. Although these methods are effective when applied to features, samples, or labels individually, they often need to be combined when dealing with multi-label data containing high-dimensional features, outlier samples, and noisy labels. However, such a combined approach tends to overlook the interactions among these dimensions, which can ultimately limit overall performance.

In addition to the aforementioned dimension-specific data reduction methods, there are few multi-label co-selection approaches that simultaneously select discriminative features and representative instances from multi-label data, such as MAVNS [Lin *et al.*, 2021] and MLVQ-JMR [Li *et al.*, 2023]. Despite the advances made by these co-selection methods,

*Corresponding Author

their effectiveness is fundamentally limited by the assumption that the observed labels are entirely accurate, which is rarely the case in practice. In real-world applications, obtaining precise label annotations for every instance is often costly and may even be infeasible. As a result, annotators typically provide label sets that include the ground-truth labels as well as noisy ones [Xie *et al.*, 2025; Li *et al.*, 2022c]. The presence of noisy labels can adversely affect the performance of data reduction methods. This can be validated by Fig. 1, which compares the performance of our proposed method EMMS with two combined data reduction approaches SILS and MALS. SILS integrates three data reduction methods for multi-label data: feature selection (SPLDG [Zhang *et al.*, 2025]), instance selection (CHC-IS [Del Castillo *et al.*, 2021]), and label selection (Inf-LS [Pan *et al.*, 2022]). MALS combines a multi-label feature and instance co-selection method (MAVNS [Lin *et al.*, 2021]) with the same label selection method Inf-LS. It can be observed that the proposed EMMS significantly outperforms the combined methods under different noisy label ratios. These results demonstrate that simple combination methods are hardly effective for multi-label data containing high-dimensional features, outlier instances, and noisy labels.

To address the aforementioned issues, we propose a novel multi-label data reduction method called Evidential Multi-label Multi-dimensional Selection (EMMS) that simultaneously selects features, instances, and labels. Specifically, EMMS introduces a dual projection mechanism with sparsity constraints that first maps high-dimensional data into a latent space and then into the label space. Meanwhile, the projection residuals are explicitly modeled to facilitate the identification of representative instances, thereby enabling the joint selection of features, instances, and labels. Furthermore, we utilize Dempster–Shafer theory [Dempster, 2008] to assign belief masses, assessing the reliability of learned labels based on both instance-level and label-level evidence. These belief masses are then aggregated to refine the learned labels and mitigate the impact of noisy labels, which in turn leads to more effective multi-dimensional selection. The overall framework of the proposed EMMS is illustrated in Fig. 2. The main contributions of this paper are as follows:

- To the best of our knowledge, this is the first work to propose and investigate the joint selection of features, instances, and labels from high-dimensional multi-label data with noisy instances and labels, thereby advancing data reduction towards more realistic scenarios.
- An evidence-based label refinement mechanism is proposed that integrates two complementary sources of evidence to enhance the reliability of learned labels and thereby facilitate multi-dimensional selection.
- An effective alternative optimization algorithm is developed for the proposed EMMS. Comprehensive experiments demonstrate that EMMS consistently outperforms state-of-the-art approaches.

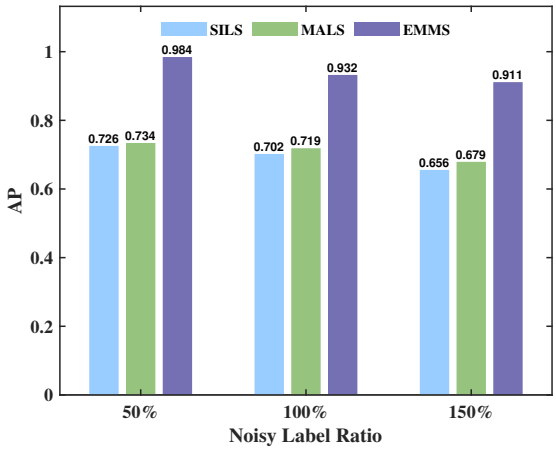

Figure 1: Performance comparison of different data reduction methods under various noisy label ratios on the Arts dataset, where the noisy label ratio is defined as the ratio of noisy labels to ground-truth labels. SILS and MALS are two combined data reduction methods.

2 The Proposed Method

2.1 Notations and Problem Definition

Throughout this paper, matrices and vectors are denoted by bold uppercase and lowercase letters, respectively. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, its trace and transpose are denoted by $\mathrm{Tr}\,(\boldsymbol{A})$ and $\boldsymbol{A}^T$. The Frobenius norm of $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}|a_{ij}|^2}$, and the ℓ_1-norm as $\|\boldsymbol{A}\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{p}|a_{ij}|$, where a_{ij} denotes the (i,j)-th entry of $\boldsymbol{A}$. $\mathbf{1}_{n\times c}$ represents an $n \times c$ matrix whose entries are all ones.

To clearly define the problem, consider a multi-label dataset $\mathcal{D} = \{(\boldsymbol{X}, \hat{\boldsymbol{Y}})\}$, where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ denotes the feature matrix containing n instances and d features, and $\hat{\boldsymbol{Y}} \in \mathbb{R}^{n\times l}$ is the observed label matrix corrupted by noise, with l being the number of label classes. In $\hat{\boldsymbol{Y}}$, $\hat{y}_{ij} = 1$ indicates that the j-th label is either relevant to instance $\boldsymbol{x}_{i\cdot}$ or is a noisy label, whereas $y_{ij} = 0$ indicates that the j-th label is irrelevant to $\boldsymbol{x}_{i\cdot}$. This study aims to simultaneously select the b most informative features, g representative instances, and h reliable labels from $\mathcal{D}$.

2.2 Dual-Projection-based Multi-Dimensional Selection for Multi-Label Data

Conventional multi-label feature selection methods [Nie *et al.*, 2010; Lei *et al.*, 2023] typically learn a direct mapping from features to labels by incorporating a regularization term, as illustrated below:

$$\min_{\boldsymbol{W}} \|\boldsymbol{X}\boldsymbol{W} - \hat{\boldsymbol{Y}}\|_F^2 + \Omega(\boldsymbol{W}), \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{d\times l}$ denotes the feature projection matrix and Ω is the regularization term that controls model complexity. Despite their effectiveness in simple settings, such formulations face two main limitations in real-world scenarios. First, data often contain noisy and outlier samples, which can dominate the reconstruction loss and bias the learned projection toward non-representative samples, resulting in unreliable feature selection. Second, the high cost and subjectivity involved

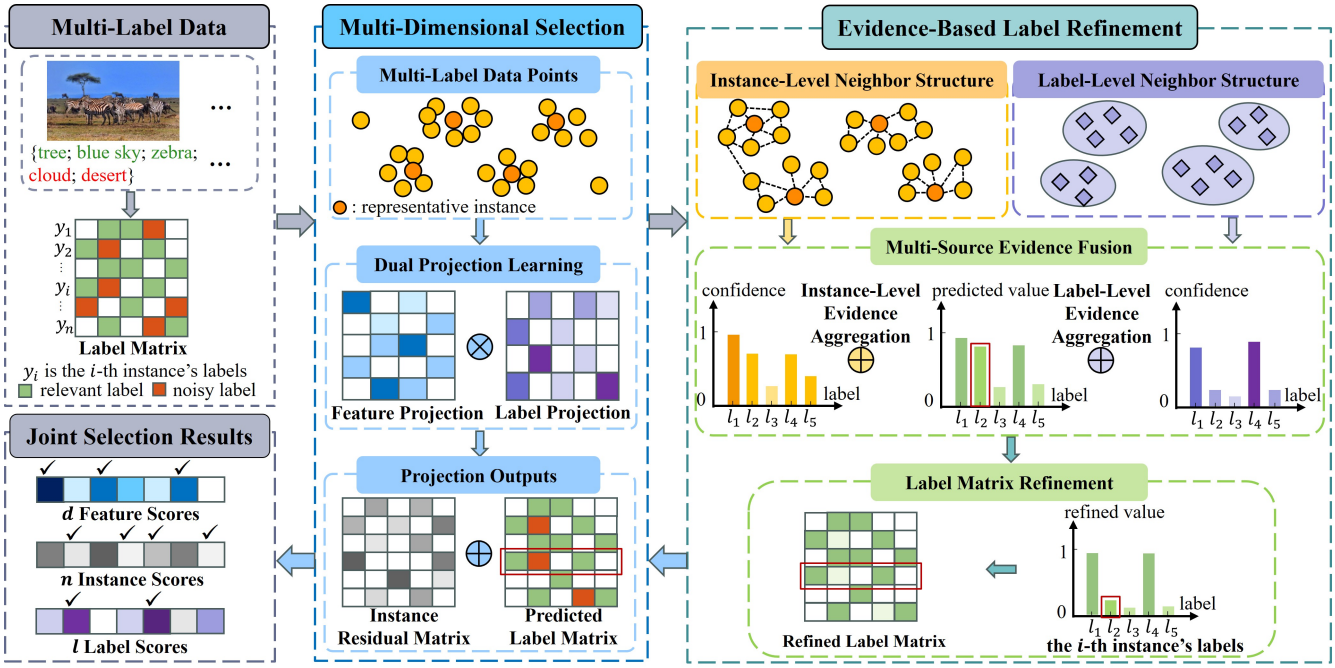

Figure 2: The framework of the proposed EMMS.

in obtaining label annotations often result in noisy labels, as limited resources and subjective judgments can introduce errors into the annotation process. Directly learning from these labels may introduce misleading supervision signals, distorting the projection matrix learning process and ultimately degrading the performance of downstream tasks.

To address the above issues, we propose a unified multi-dimensional selection framework that introduces a sparsity-regularized dual projection mechanism to map high-dimensional features first into a latent space and subsequently into the label space. In addition, projection residuals are explicitly modeled to identify representative instances, enabling the joint selection of features, instances, and labels. This can be formulated as follows:

$$\begin{aligned}\min_{\substack{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{R},\\ \boldsymbol{Y},\boldsymbol{N}}} & \frac{1}{2}\|\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^\top - \boldsymbol{R} - \boldsymbol{Y}\|_F^2 + \lambda_1\|\boldsymbol{P}\|_1 + \lambda_2\|\boldsymbol{Q}\|_1 \\ & + \lambda_3(\|\boldsymbol{R}\|_1 + \|\boldsymbol{N}\|_1) \\ s.t.\ & \boldsymbol{P},\boldsymbol{Q},\boldsymbol{R},\boldsymbol{Y},\boldsymbol{N} \geq 0, \hat{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{N},\end{aligned} \tag{2}$$

where $\boldsymbol{P} \in \mathbb{R}^{d\times c}$ is the feature projection matrix that maps the original high-dimensional features into a latent space, and $\boldsymbol{Q}^\top \in \mathbb{R}^{c\times l}$ serves as a projection from the latent feature space to the label space. Besides, $\boldsymbol{R}$ is a projection residual matrix that captures instance-level deviations, while $\boldsymbol{Y}$ and $\boldsymbol{N}$ denote the learned ground-truth label matrix and a sparse, non-negative noise matrix, respectively, such that $\hat{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{N}$. The ℓ_1-norm regularization terms imposed on $\boldsymbol{P}$, $\boldsymbol{Q}$, $\boldsymbol{R}$, and $\boldsymbol{N}$ promote sparsity, thereby suppressing irrelevant features, non-representative instances, and noisy labels, respectively, within a unified framework. Meanwhile, λ_1, λ_2, and λ_3 are three regularization parameters. Then, based on Eq. (2), we can select the top b features, g instances, and h labels by ranking the row-wise ℓ_2-norms in $\boldsymbol{P}$, $\boldsymbol{R}$, and $\boldsymbol{Q}$ in descending, ascending, and descending order, respectively. Hence, Eq. (2) utilizes a dual projection mechanism with sparse regularization to enable the simultaneous selection of features, instances, and labels.

Furthermore, we integrate two manifold regularization terms into Eq. (2) to preserve the local neighborhood structures in both the feature and label spaces, leading to the following formulation:

$$\begin{aligned}\min_{\substack{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{R},\\ \boldsymbol{Y},\boldsymbol{N}}} & \frac{1}{2}\|\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^\top - \boldsymbol{R} - \boldsymbol{Y}\|_F^2 + \lambda_1\|\boldsymbol{P}\|_1 + \lambda_2\|\boldsymbol{Q}\|_1 \\ & + \lambda_3(\|\boldsymbol{R}\|_1 + \|\boldsymbol{N}\|_1) + \frac{1}{2}\mathrm{Tr}(\boldsymbol{P}^\top\boldsymbol{L}_F\boldsymbol{P}) + \frac{1}{2}\mathrm{Tr}(\boldsymbol{Y}^\top\boldsymbol{L}_X\boldsymbol{Y}) \\ s.t.\ & \boldsymbol{P},\boldsymbol{Q},\boldsymbol{R},\boldsymbol{Y},\boldsymbol{N} \geq 0, \hat{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{N},\end{aligned} \tag{3}$$

where $\boldsymbol{L}_F = \boldsymbol{D}_F - \boldsymbol{S}_F$ and $\boldsymbol{L}_X = \boldsymbol{D}_X - \boldsymbol{S}_X$ denote the Laplacian matrices constructed from the cosine similarity matrices $\boldsymbol{S}_F \in \mathbb{R}^{d\times d}$ and $\boldsymbol{S}_X \in \mathbb{R}^{n\times n}$ over features and samples, respectively. The corresponding degree matrices $\boldsymbol{D}_F$ and $\boldsymbol{D}_X$ have diagonal entries $\sum_{j=1}^{d}(\boldsymbol{S}_F)_{ij}$ and $\sum_{j=1}^{n}(\boldsymbol{S}_X)_{ij}$, respectively. The last two terms in Eq. (3) encourage similar features to stay close in the latent space and correlated instances to share similar labels. This facilitates the learning of more accurate projections and labels, thereby enhancing the effectiveness of joint feature, instance, and label selection [Li *et al.*, 2022a; Braytee *et al.*, 2017].

2.3 Evidence-based Label Refinement

Label manifold regularization encourages neighboring instances to have similar label predictions. However, if an in-

stance is assigned incorrect labels, this constraint can propagate noisy label information to its neighbors, resulting in an unreliable learned label matrix $\boldsymbol{Y}$, which in turn reduces performance on downstream tasks.

To address this issue, we incorporate Dempster–Shafer theory (DST) [Denœux, 2019; Li *et al.*, 2022b] to refine the learned label matrix $\boldsymbol{Y}$ in a reliability-aware manner. By associating each label assignment with a belief mass and refining unreliable predictions through the fusion of multiple sources of evidence, our approach provides more reliable supervision for multi-dimensional selection. Specifically, since each entry in $\boldsymbol{Y}$ corresponds to a binary label, learning $\boldsymbol{Y}$ can naturally be decomposed into a set of binary label assignment problems, which allows DST to be applied in a label-wise manner. Hence, for each label assignment y_{ij}, we define a binary frame of discernment $\Theta = \{+, -\}$, where "+" and "−" denote the hypotheses that the j-th label is relevant or irrelevant to the i-th instance, respectively. The corresponding power set is given by $2^{\Theta} = \{\emptyset, \{+\}, \{-\}, \{+, -\}\}$, where the universal set $\{+, -\}$ represents uncertainty regarding the current label assignment. Then, a basic probability assignment (BPA) is constructed for each y_{ij}, which allocates belief masses to the positive, negative, and uncertain hypotheses as follows:

$$\begin{cases} m_{ij}(\{+\}) = c_{ij} \cdot y_{ij}, \\ m_{ij}(\{-\}) = c_{ij} \cdot (1 - y_{ij}), \\ m_{ij}(\Theta) = 1 - c_{ij}, \end{cases} \tag{4}$$

where $m(\cdot)$ is a mapping from 2^{Θ} to $[0, 1]$, satisfying $m(\emptyset) = 0$ and $\sum_{A \in 2^{\Theta}} m(A) = 1$. Thus, $m(A)$ represents the belief mass assigned to the event A, where $A \in 2^{\Theta}$. $c_{ij} \in [0, 1]$ denotes the confidence in the current label assignment, where higher confidence concentrates more belief on specific hypotheses, while lower confidence results in more mass being assigned to uncertainty.

To comprehensively assess the confidence of the current label assignment y_{ij}, we leverage two complementary sources of evidence derived from instance-level and label-level structural information, based on the following assumptions. First, from the instance perspective, if instances similar to the i-th instance are assigned the j-th label, the assignment of y_{ij} as positive is considered more reliable. Second, from the label perspective, if labels correlated with the j-th label also co-occur in the i-th instance, assigning label j is deemed more credible. Hence, the instance correlations are captured by the similarity matrix $\boldsymbol{S}_X$, while the label correlations $\boldsymbol{S}_Y \in \mathbb{R}^{l \times l}$ are computed as the cosine similarity between label representations derived from $\boldsymbol{Q}$. Based on these two sources of evidence, the confidence for y_{ij} is calculated as follows:

$$c_{ij}^{(k)} = \exp\left(-\frac{(y_{ij} - e_{ij}^{(k)})^2}{\sigma^{(k)}}\right), \quad k = 1, 2, \tag{5}$$

where $e_{ij}^{(1)} = \boldsymbol{S}_X(i, \cdot)\boldsymbol{y}_{\cdot j}$ and $e_{ij}^{(2)} = \boldsymbol{y}_{i\cdot}\boldsymbol{S}_Y(\cdot, j)$ denote the instance-level and label-level evidence, respectively, and $\sigma^{(k)}$ controls the sensitivity of confidence estimation.

Based on the confidence values $c_{ij}^{(k)}(k = 1, 2)$, two corresponding BPAs $m_{ij}^{(k)}(k = 1, 2)$ are constructed using Eq. (4), each encoding the belief masses for the positive, negative, and uncertain hypotheses from different perspectives. These two BPAs are then fused to aggregate consistent evidence and normalize conflicting information, using the following combination rules:

$$\begin{cases} m_{ij}^{(f)}(\{+\}) = \frac{1}{1-\mathcal{K}_{ij}}(m_{ij}^{(1)}(\{+\})m_{ij}^{(2)}(\{+\}) + \\ \qquad m_{ij}^{(1)}(\{+\})m_{ij}^{(2)}(\Theta) + m_{ij}^{(1)}(\Theta)m_{ij}^{(2)}(\{+\})), \\ m_{ij}^{(f)}(\{-\}) = \frac{1}{1-\mathcal{K}_{ij}}(m_{ij}^{(1)}(\{-\})m_{ij}^{(2)}(\{-\}) + \\ \qquad m_{ij}^{(1)}(\{-\})m_{ij}^{(2)}(\Theta) + m_{ij}^{(1)}(\Theta)m_{ij}^{(2)}(\{-\})), \\ m_{ij}^{(f)}(\Theta) = \frac{1}{1-\mathcal{K}_{ij}}m_{ij}^{(1)}(\Theta)m_{ij}^{(2)}(\Theta), \end{cases} \tag{6}$$

where $m_{ij}^{(f)}$ denotes the aggregated belief mass, and $\mathcal{K}_{ij}$ is the conflict coefficient between the two evidence sources. Thus, the aggregated belief mass $m_{ij}^{(f)}$ provides a reliability-aware refinement of the learned label assignment, as given by:

$$y_{ij} = m_{ij}^{(f)}(\{+\}) + \tau \cdot m_{ij}^{(f)}(\Theta), \tag{7}$$

where $\tau \in [0, 1]$ controls the contribution of uncertainty. During subsequent optimization, these refined labels are iteratively updated, allowing the model to gradually suppress noisy assignments and strengthen reliable ones, thereby enhancing the overall quality of the learned label matrix $\boldsymbol{Y}$.

3 Optimization and Analyses

Since the objective function in Eq. (3) is non-convex with respect to all variables, we develop an iterative optimization algorithm that updates each variable in turn while keeping the others fixed.

Update $\boldsymbol{P}$ and fix others. With the other variables fixed, following [Ma and Chow, 2018], the objective function with respect to $\boldsymbol{P}$ can be rewritten as:

$$\min_{\boldsymbol{P} \geq 0} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^{\top} - \boldsymbol{R} - \boldsymbol{Y}\|_F^2 + \lambda_1 \mathrm{Tr}(\mathbf{1}_{d\times c}^T \boldsymbol{P}) - \frac{1}{2}\mathrm{Tr}(\boldsymbol{P}^{\top}\boldsymbol{L}_F\boldsymbol{P}). \tag{8}$$

By differentiating Eq. (8) w.r.t. $\boldsymbol{P}$, we can obtain the following update rule for $\boldsymbol{P}$:

$$p_{ij} \leftarrow p_{ij} \frac{(\boldsymbol{X}^{\top}(\boldsymbol{R} + \boldsymbol{Y})\boldsymbol{Q} + \boldsymbol{S}_F\boldsymbol{P})_{ij}}{(\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^{\top}\boldsymbol{Q} + \lambda_1 \mathbf{1}_{d\times c} + \boldsymbol{D}_F\boldsymbol{P})_{ij}}. \tag{9}$$

Update $\boldsymbol{Q}$ and fix others. When the other variables are fixed, optimizing $\boldsymbol{Q}$ in Eq. (3) can be reformulated as follows:

$$\min_{\boldsymbol{Q} \geq 0} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^{\top} - \boldsymbol{R} - \boldsymbol{Y}\|_F^2 + \lambda_2\|\boldsymbol{Q}\|_1. \tag{10}$$

Analogous to the update of $\boldsymbol{P}$, the update rule for $\boldsymbol{Q}$ can be derived as follows:

$$q_{ij} \leftarrow q_{ij} \frac{((\boldsymbol{R} + \boldsymbol{Y})^{\top}\boldsymbol{X}\boldsymbol{P})_{ij}}{(\boldsymbol{Q}\boldsymbol{P}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{P} + \lambda_2 \mathbf{1}_{l\times c})_{ij}}. \tag{11}$$

Update $\boldsymbol{R}$ and fix others. After fixing the other variables, $\boldsymbol{R}$ can be updated by solving the following optimization problem:

$$\min_{\boldsymbol{R} \geq 0} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^{\top} - \boldsymbol{R} - \boldsymbol{Y}\|_F^2 + \lambda_3\|\boldsymbol{R}\|_1. \tag{12}$$

Algorithm 1 Iterative Algorithm of EMMS

Inputs: Multi-label data $\mathcal{D} = \{(\boldsymbol{X}, \hat{\boldsymbol{Y}})\}$, and parameters λ_1, λ_2 and λ_3.

Initialize $\boldsymbol{P}$, $\boldsymbol{Q}$, $\boldsymbol{R}$, $\boldsymbol{N}$, and $\boldsymbol{Y}$.
Compute similarity matrices $\boldsymbol{S}_F$ and $\boldsymbol{S}_X$.
while not converge **do**
  Update $\boldsymbol{P}$ via Eq. (9);
  Update $\boldsymbol{Q}$ via Eq. (11);
  Update $\boldsymbol{R}$ via Eq. (13);
  Update $\boldsymbol{Y}$ via Eq. (15);
  Refine $\boldsymbol{Y}$ according to Eq.(7);
  Update $\boldsymbol{N}$ via Eq. (17);
end while

Output: The top b features, g instances, and h labels are jointly selected by ranking the row-wise ℓ_2-norms of $\boldsymbol{P}$ in descending order, $\boldsymbol{R}$ in ascending order, and $\boldsymbol{Q}$ in descending order, respectively.

Taking the derivative of Eq. (12) with respect to $\boldsymbol{R}$, we obtain the following update rule for $\boldsymbol{R}$:

$$r_{ij} \leftarrow r_{ij} \frac{(\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^\top)_{ij}}{(\boldsymbol{R} + \boldsymbol{Y} + \lambda_3 \mathbf{1}_{n\times l})_{ij}}. \tag{13}$$

Update $\boldsymbol{Y}$ and fix others. Following the same approach as for updating $\boldsymbol{P}$, we derive the optimization problem for $\boldsymbol{Y}$ as follows:

$$\min_{\boldsymbol{Y}\geq 0, \hat{\boldsymbol{Y}}=\boldsymbol{Y}+\boldsymbol{N}} \|\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^\top - \boldsymbol{R} - \boldsymbol{Y}\|_F^2 + \mathrm{Tr}(\boldsymbol{Y}^\top \boldsymbol{L}_X \boldsymbol{Y}) \tag{14}$$

Then, the update rule for $\boldsymbol{Y}$ is given by:

$$y_{ij} \leftarrow y_{ij} \frac{(\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^\top + \mu\hat{\boldsymbol{Y}} + \boldsymbol{S}_X\boldsymbol{Y})_{ij}}{(\boldsymbol{Y} + \mu(\boldsymbol{Y} + \boldsymbol{N}) + \boldsymbol{R} + \boldsymbol{D}_X\boldsymbol{Y})_{ij}}. \tag{15}$$

After updating $\boldsymbol{Y}$ according to Eq. (15), its elements are subsequently refined using Eq. (7).

Update $\boldsymbol{N}$ and fix others. By fixing the other variables, the optimization problem for $\boldsymbol{N}$ is reformulated as follows:

$$\min_{\boldsymbol{N}\geq 0, \hat{\boldsymbol{Y}}=\boldsymbol{Y}+\boldsymbol{N}} \lambda_3 \|\boldsymbol{N}\|_1. \tag{16}$$

Using the same approach as for updating $\boldsymbol{P}$, $\boldsymbol{N}$ can be updated according to the following rule:

$$n_{ij} \leftarrow n_{ij} \frac{(\mu\hat{\boldsymbol{Y}})_{ij}}{(\mu(\boldsymbol{Y} + \boldsymbol{N}) + \lambda_3 \mathbf{1}_{n\times l})_{ij}}. \tag{17}$$

Algorithm 1 summarizes the overall optimization procedure of the proposed EMMS. In this algorithm, $\boldsymbol{P}$, $\boldsymbol{Q}$, $\boldsymbol{R}$, and $\boldsymbol{N}$ are randomly initialized to non-negative matrices, while $\boldsymbol{Y}$ is initialized to $\hat{\boldsymbol{Y}}$.

Complexity and Convergence Analysis. The computational complexity of Algorithm 1 is mainly due to matrix multiplication operations. Specifically, updating $\boldsymbol{P}$ and $\boldsymbol{Q}$ each requires complexity of $\mathcal{O}(nd^2)$. For updating $\boldsymbol{R}$, the time complexity is $\mathcal{O}(ndc)$. Updating $\boldsymbol{Y}$ has a complexity of $\mathcal{O}(ndc + n^2 l)$, and its refinement step takes $\mathcal{O}(n^2 l)$ complexity. The update of $\boldsymbol{N}$ consists solely of element-wise operations, whose computational cost can be ignored. Therefore, the overall computational complexity of Algorithm 1 is $\mathcal{O}(nd(c+d)+n^2 l)$. Due to space limitations, the convergence proof of Algorithm 1 is provided in the supplementary material[1].

Datasets	Instances	Features	Labels	Card.	Dens.
Staexp	3971	842	233	2.272	0.010
Corel5k	5000	499	374	3.522	0.009
Arts	5000	462	26	1.636	0.063
Science	5000	743	40	1.451	0.036
Social	5000	1047	39	1.283	0.033
Co16k1	13770	500	153	2.859	0.019

Table 1: Dataset description

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on six real-world multi-label datasets: Stackex_philosophy[2] (Staexp), Corel5k[2], Arts[3], Science[3], Social[3], and Corel16k001[2] (Co16k1). Details of these datasets are summarized in Table 1, where "Card." denotes the average number of labels per instance and "Dens." represents the label density of each dataset.

Compared Methods. To evaluate the effectiveness of the proposed EMMS, we compared it with several state-of-the-art (SOTA) methods. These include the multi-label triple-selection method mFILS [Mansouri and Benabdeslem, 2021], as well as nine combination-based methods for multi-label data that integrate feature selection, instance selection, and label selection: SILS (SPLDG [Zhang *et al.*, 2025] with CHC-IS [Del Castillo *et al.*, 2021] and Inf-LS [Pan *et al.*, 2022]); NILS (NMMFS [Wang *et al.*, 2025] with CHC-IS and Inf-LS); PILS (PML-FSLA [Pan *et al.*, 2025] with CHC-IS and Inf-LS); RILS (ROAD [Zhang *et al.*, 2023] with CHC-IS and Inf-LS); IFLS (IS-FS [Kusy and Zajdel, 2024] with Inf-LS); FILS (FS-IS [Kusy and Zajdel, 2024] with Inf-LS); MJLS (MLVQ-JMR [Li *et al.*, 2023] with Inf-LS); s2LS (sCOs2 [Benabdeslem *et al.*, 2022] with Inf-LS); and MALS (MAVNS [Lin *et al.*, 2021] with Inf-LS).

Comparison Schemes. For a fair comparison, all methods are tuned using grid search, and their best results are reported. The parameters λ_1, λ_2, and λ_3 in our method are selected from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$, while those for the compared methods are set as recommended in their respective papers. Since it is challenging to determine the optimal number of selected features, instances, and labels [Siblini *et al.*, 2019; Ros and Guillaume, 2019], we set their selection ratios to range from 10% to 50% in increments of 10%. Consistent with previous methods [Sun *et al.*, 2019; Xie and Huang, 2021], false-positive labels are randomly introduced for each instance at a rate equal to 50% of the number of relevant labels to simulate a noisy environment. Fol-

[1] https://github.com/yangdali706/SupplementaryMaterial
[2] http://www.uco.es/kdis/mllresources
[3] http://www.lamda.nju.edu.cn/code_MDDM.ashx

Methods	AP (the larger the better)						RL (the smaller the better)					
	Staexp	Corel5k	Arts	Science	Social	Co16k1	Staexp	Corel5k	Arts	Science	Social	Co16k1
EMMS	**0.6056**	**0.6259**	**0.9845**	**0.9313**	**0.9700**	**0.7262**	**0.1306**	**0.0792**	**0.0172**	**0.0579**	**0.0294**	**0.1487**
mFILS	0.4878	0.3799	0.8264	0.7650	0.7337	0.3595	0.1926	0.2083	0.2174	0.2354	0.2477	0.3277
s2LS	0.3247	0.3936	0.6709	0.7226	0.7318	0.4155	0.2952	0.1508	0.4374	0.2666	0.2644	0.3178
MALS	0.3165	0.3970	0.7340	0.6590	0.7384	0.4241	0.2941	0.1493	0.4044	0.3259	0.2638	0.3036
MJLS	0.3098	0.3923	0.7024	0.6876	0.7889	0.4104	0.3073	0.1479	0.4088	0.2961	0.2081	0.3088
IFLS	0.3456	0.3774	0.6998	0.7142	0.7681	0.4130	0.2738	0.1564	0.3983	0.2612	0.2311	0.3068
FILS	0.3221	0.2612	0.7168	0.6957	0.8038	0.3922	0.2936	0.2041	0.3705	0.2855	0.1974	0.3222
RILS	0.5004	0.3525	0.6783	0.6415	0.7559	0.4232	0.1507	0.1964	0.4217	0.3521	0.2432	0.2989
SILS	0.3619	0.4030	0.7255	0.6814	0.7777	0.4213	0.2702	0.1456	0.4131	0.3009	0.2286	0.3009
NILS	0.3008	0.3989	0.6948	0.6829	0.7762	0.4156	0.2970	0.1494	0.4051	0.3034	0.2234	0.3048
PILS	0.3258	0.4008	0.7145	0.6803	0.7843	0.4193	0.2810	0.1443	0.3836	0.3126	0.2215	0.3041

Table 2: Performance comparison between EMMS and other methods on six datasets in terms of AP and RL.

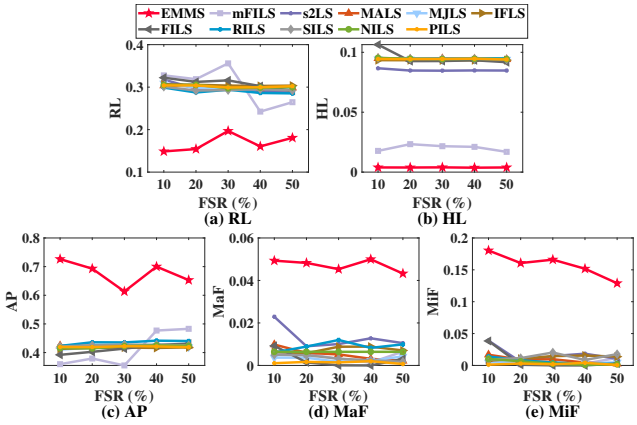

Figure 3: Comparison of different methods on Co16k1 dataset across varying feature selection ratios using five metrics.

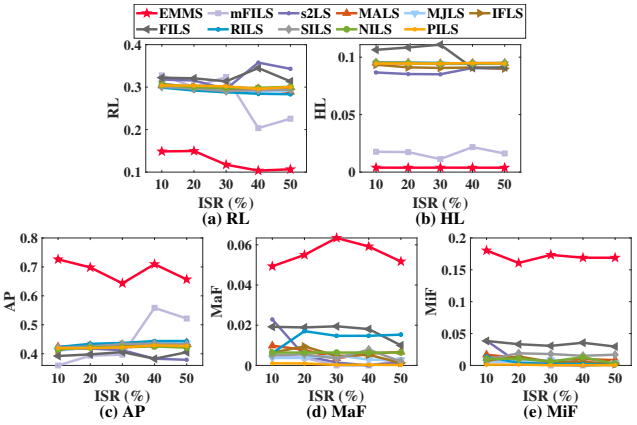

Figure 4: Comparison of different methods on Co16k1 dataset across varying instance selection ratios using five metrics.

lowing [Mansouri and Benabdeslem, 2021], we first apply the proposed methods to identify the most informative features, instances, and labels across all datasets. Each dataset is then represented using these selected features and labels. An ML-KNN classifier [Zhang and Zhou, 2007] is trained on the selected instances and their corresponding labels, and is subsequently used to predict the labels of the remaining samples. Five widely used metrics, including Ranking Loss (RL), Hamming Loss (HL), Average Precision (AP), Macro-F1 (MaF), and Micro-F1 (MiF), are used to evaluate the results, with lower HL and RL and higher AP, MaF, and MiF indicating better performance. Each experiment is repeated five times, and the average values are reported.

4.2 Experimental Results and Analysis

Performance Comparison. Table 2 presents the performance of different methods in terms of AP and RL on six datasets, with 10% of the instances, features, and labels are selected. The best results are highlighted in bold, while the second-best results are underlined. Due to limited space, the results for MaF, MiF, and HL are included in the supplementary material. As shown in Table 2, EMMS consistently outperforms all other competing methods across all

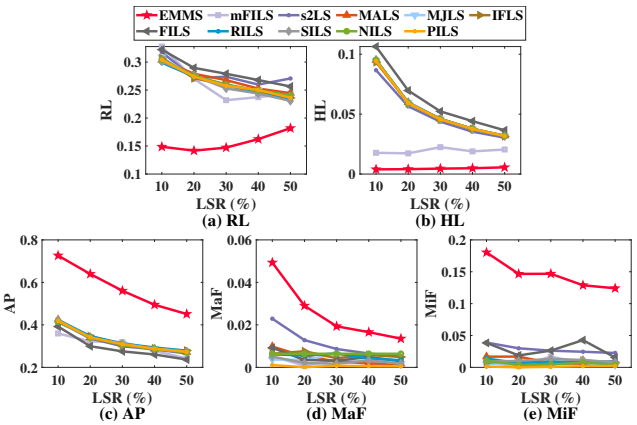

Figure 5: Comparison of different methods on Co16k1 dataset across varying label selection ratios using five metrics.

datasets. EMMS achieves more than a 20% improvement in AP over the second-best method on Corel5k and Co16k1 datasets, and at least a 10% improvement on the remaining datasets. Additionally, EMMS reduces RL by more than 15%

Methods	Staexp	Corel5k	Arts	Science	Social	Co16k1
EMMS	**0.6056**	**0.6259**	**0.9845**	**0.9313**	**0.9700**	**0.7262**
EMMS-I	0.5192	0.3519	0.7448	0.9031	0.8845	0.4430
EMMS-II	0.2366	0.2308	0.7590	0.6436	0.8387	0.2859

Table 3: Ablation results of EMMS on six datasets in terms of AP.

on Arts, Science, Social, and Co16k1 datasets compared to the second-best methods. Furthermore, supplementary results from the Friedman and Nemenyi post-hoc tests demonstrate that EMMS significantly outperforms other methods.

Furthermore, to provide a comprehensive evaluation of EMMS, we also present the results for all methods across all datasets using five metrics, with varying ratios of feature, instance, and label selection. Figs. 3, 4, and 5 show the results for the Co16k1 dataset with different feature selection ratios (FSR), instance selection ratios (ISR), and label selection ratios (LSR), respectively. The results for the other five datasets are available in the supplementary material. As illustrated in these figures, the proposed EMMS outperforms the other methods in most cases. The superior performance of EMMS can be attributed to the joint selection of features, instances, and labels, combined with evidence-based label refinement, as these components mutually reinforce each other.

Ablation Study. We conduct an ablation study to evaluate the contributions of the key modules in EMMS. Table 3 reports the AP results, where EMMS-I and EMMS-II are the variants without the evidence-based label refinement module and the manifold learning module, respectively. As shown in Table 3, EMMS outperforms both variants, confirming the effectiveness of the two modules.

Figure 6: An image from the Scene dataset includes six candidate labels: "beach", "mountain", "fall foliage", "field", "sunset", and "urban". In this image, "beach" and "mountain" are ground-truth labels, while "fall foliage" and "field" are noisy labels.

Case Study. We conduct an experiment on the Scene dataset [Boutell *et al.*, 2004] with noisy labels to evaluate EMMS's ability to identify relevant labels. Fig. 6 shows an image from the Scene dataset, featuring two ground-truth labels (green) and two noisy labels (red). Table 4 presents the label selection results for different methods. As shown, EMMS identifies relevant labels more accurately than other methods, which tend to select noisy or irrelevant labels. This demonstrates the effectiveness of EMMS in handling multi-label data with noisy labels.

Parameter Sensitivity and Convergence Analysis. The proposed EMMS includes three tuning parameters, namely λ_1, λ_2, and λ_3. Fig. 7 illustrates the sensitivity of the AP metric on Science dataset under different parameters, showing that the performance of EMMS is relatively stable with respect to λ_1 and λ_3, and generally performs better with smaller values of λ_2. Additionally, Fig. 8 presents the convergence curves of EMMS on Staexp, Arts, and Co16k1 datasets. As shown in the figure, EMMS decreases rapidly during the initial iterations and stabilizes after about 10 iterations.

Methods	beach	mountain	fall foliage	field	sunset	urban
EMMS	✓	✓				
mFILS		✓		✓		
s2LS	✓			✓		
MALS				✓	✓	
MJLS	✓		✓			
IFLS		✓		✓		
FILS			✓	✓		
RILS				✓	✓	
SILS		✓	✓			
NILS	✓		✓			
PILS			✓	✓		

Table 4: Selected labels by different methods.

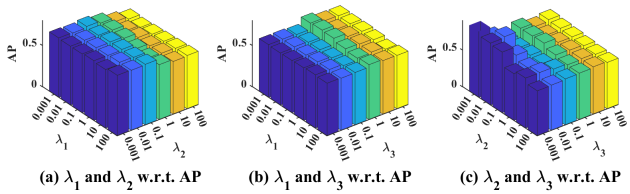

Figure 7: AP of EMMS with varying parameters λ_1, λ_2, and λ_3 on Science dataset.

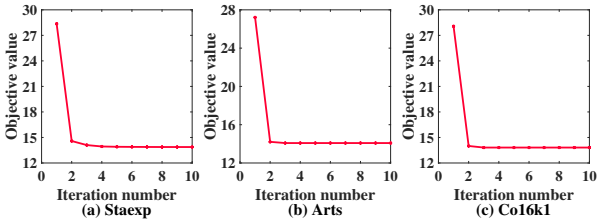

Figure 8: Convergence curves of EMMS on Staexp, Arts, and Co16k1 datasets.

5 Conclusion

In this paper, we propose a novel multi-label data reduction method called EMMS to simultaneously select features, instances, and labels from high-dimensional multi-label data with noisy labels. Unlike existing approaches that treat feature, sample, and label selection separately, we integrate them into a unified framework through a dual projection mechanism with sparsity constraints. Meanwhile, we employ an evidence-based label refinement module to improve the reliability of learned labels and reduce the influence of noisy labels, thereby enhancing the effectiveness of joint selection. Extensive experiments demonstrate the superiority of EMMS compared to SOTA methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 72495122), and the Natural Science Foundation Project of Sichuan Province (No. 2024NSFSC0504). The author Yanyong Huang gratefully acknowledges the support of K. C. Wong Education Foundation and DAAD.

References

[Benabdeslem *et al.*, 2022] Khalid Benabdeslem, Dou El Kefel Mansouri, and Raywat Makkhongkaew. sCOs: Semi-supervised co-selection by a similarity preserving approach. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2899–2911, 2022.

[Boutell *et al.*, 2004] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[Braytee *et al.*, 2017] Ali Braytee, Wei Liu, Daniel R Catchpoole, and Paul J Kennedy. Multi-label feature selection using correlation information. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1649–1656, 2017.

[Del Castillo *et al.*, 2021] Juan Antonio Romero Del Castillo, Domingo Ortiz-Boyer, and Nicolás García-Pedrajas. Instance selection for multi-label learning based on a scalable evolutionary algorithm. In *Proceedings of the 2021 International Conference on Data Mining Workshops*, pages 843–851. IEEE, 2021.

[Dempster, 2008] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 57–72. Springer, 2008.

[Denœux, 2019] Thierry Denœux. Logistic regression, neural networks and dempster–shafer theory: A new perspective. *Knowledge-Based Systems*, 176:54–67, 2019.

[Kusy and Zajdel, 2024] Maciej Kusy and Roman Zajdel. New data reduction algorithms based on the fusion of instance and feature selection. *Knowledge-Based Systems*, 296:111844, 2024.

[Lei *et al.*, 2023] Yu Lei, Qin Li, and Jane You. Efficient regression with feature selection based on l12-norm. In *Proceedings of the 2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning*, pages 250–258. IEEE, 2023.

[Li *et al.*, 2022a] Junlong Li, Peipei Li, Xuegang Hu, and Kui Yu. Learning common and label-specific features for multi-label classification with correlation information. *Pattern Recognition*, 121:108259, 2022.

[Li *et al.*, 2022b] Qing Li, Changqing Zhang, Qinghua Hu, Huazhu Fu, and Pengfei Zhu. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 25:3420–3431, 2022.

[Li *et al.*, 2022c] Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label correlations for noisy multi-label learning. *Advances in Neural Information Processing Systems*, 35:24184–24198, 2022.

[Li *et al.*, 2023] Haikun Li, Min Fang, and Peng Wang. Dual dimensionality reduction on instance-level and feature-level for multi-label data. *Neural Computing and Applications*, 35(35):24773–24782, 2023.

[Li *et al.*, 2024] Haikun Li, Min Fang, Hang Li, and Peng Wang. Prototype selection for multi-label data based on label correlation. *Neural Computing and Applications*, 36(5):2121–2130, 2024.

[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.

[Lin *et al.*, 2021] Chun-Cheng Lin, Jia-Rong Kang, Yu-Lin Liang, and Chih-Chi Kuo. Simultaneous feature and instance selection in big noisy data using memetic variable neighborhood search. *Applied Soft Computing*, 112:107855, 2021.

[Ma and Chow, 2018] Jianghong Ma and Tommy WS Chow. Robust non-negative sparse graph for semi-supervised multi-label learning with missing labels. *Information Sciences*, 422:336–351, 2018.

[Mansouri and Benabdeslem, 2021] Dou El Kefel Mansouri and Khalid Benabdeslem. Towards multi-label feature selection by instance and label selections. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 233–244. Springer, 2021.

[Nguyen *et al.*, 2024] Bach Hoai Nguyen, Bing Xue, and Mengjie Zhang. Evolutionary label selection for multi-label classification. In *Proceedings of the 2024 IEEE Congress on Evolutionary Computation*, pages 01–08. IEEE, 2024.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint l2, 1-norms minimization. *Advances in Neural Information Processing Systems*, 23, 2010.

[Ougiaroglou *et al.*, 2023] Stefanos Ougiaroglou, Panagiotis Filippakis, Georgia Fotiadou, and Georgios Evangelidis. Data reduction via multi-label prototype generation. *Neurocomputing*, 526:1–8, 2023.

[Pan *et al.*, 2022] Yuchen Pan, Jun Li, and Jianhua Xu. Infinite label selection method for mutil-label classification. In *Proceedings of the International Conference on Neural Information Processing*, pages 361–372. Springer, 2022.

[Pan *et al.*, 2025] Hanlin Pan, Kunpeng Liu, and Wanfu Gao. Reconsidering feature structure information and latent space alignment in partial multi-label feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19786–19794, 2025.

[Ros and Guillaume, 2019] Frédéric Ros and Serge Guillaume. From supervised instance and feature selection

algorithms to dual selection: A review. *Sampling Techniques for Supervised or Unsupervised Tasks*, pages 83–128, 2019.

[Siblini *et al.*, 2019] Wissam Siblini, Pascale Kuntz, and Frank Meyer. A review on dimensionality reduction for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 33(3):839–857, 2019.

[Sun *et al.*, 2019] Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5016–5023, 2019.

[Wang *et al.*, 2016] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016.

[Wang *et al.*, 2025] Yan Wang, Changzhong Wang, Tingquan Deng, and Wenqi Li. Multi-label feature selection via nonlinear mapping and manifold regularization. *Information Sciences*, 704:121965, 2025.

[Xie and Huang, 2021] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3676–3687, 2021.

[Xie *et al.*, 2025] Hao Xie, Ivy Liu, Bing Xue, and Mengjie Zhang. Partial multi-label feature selection via adaptive dual-graph regularization. *Knowledge-Based Systems*, 326:114077, 2025.

[Zhang and Zhou, 2006] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

[Zhang and Zhou, 2007] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[Zhang *et al.*, 2023] Zan Zhang, Zhe Zhang, Jialu Yao, Lin Liu, Jiuyong Li, Gongqing Wu, and Xindong Wu. Multi-label feature selection via adaptive label correlation estimation. *ACM Transactions on Knowledge Discovery from Data*, 17(9):1–28, 2023.

[Zhang *et al.*, 2025] Yao Zhang, Jun Tang, Ziqiang Cao, and Han Chen. Sparse multi-label feature selection via pseudo-label learning and dynamic graph constraints. *Information Fusion*, 118:102975, 2025.