applied sciences

Article

Robust Ranking Kernel Support Vector Machine via Manifold Regularized Matrix Factorization for Multi-Label Classification

Heping Song [1,2], Yiming Zhou [1], Ebenezer Quayson [1,3], Qian Zhu [1] and Xiangjun Shen [1,*]

[1] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China; songhp@ujs.edu.cn (H.S.)
[2] Jiangsu Engineering Research Center of Big Data Ubiquitous Perception and Intelligent Agriculture Applications, Zhenjiang 212013, China
[3] Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani 00233, Ghana
[*] Correspondence: xjshen@ujs.edu.cn

Abstract: Multi-label classification has been extensively researched and utilized for several decades. However, the performance of these methods is highly susceptible to the presence of noisy data samples, resulting in a significant decrease in accuracy when noise levels are high. To address this issue, we propose a robust ranking support vector machine (Rank-SVM) method that incorporates manifold regularized matrix factorization. Unlike traditional Rank-SVM methods, our approach integrates feature selection and multi-label learning into a unified framework. Within this framework, we employ matrix factorization to learn a low-rank robust subspace within the input space, thereby enhancing the robustness of data representation in high-noise conditions. Additionally, we incorporate manifold structure regularization into the framework to preserve manifold relationships among low-rank samples, which further improves the robustness of the low-rank representation. Leveraging on this robust low-rank representation, we extract a resilient low-rank features and employ them to construct a more effective classifier. Finally, the proposed framework is extended to derive a kernelized ranking approach, for the creation of nonlinear multi-label classifiers. To effectively solve this non-convex kernelized method, we employ the augmented Lagrangian multiplier (ALM) and alternating direction method of multipliers (ADMM) techniques to obtain the optimal solution. Experimental evaluations conducted on various datasets demonstrate that our framework achieves superior classification results and significantly enhances performance in high-noise scenarios.

Keywords: multi-label classification; low rank; ranking support vector machine; manifold regularized matrix factorization; unified framework

Citation: Song, H.; Zhou, Y.; Quayson, E.; Zhu, Q.; Shen, X. Robust Ranking Kernel Support Vector Machine via Manifold Regularized Matrix Factorization for Multi-Label Classification. *Appl. Sci.* **2024**, *14*, 638. https://doi.org/10.3390/app14020638

Academic Editor: Douglas O'Shaughnessy

Received: 18 October 2023
Revised: 5 January 2024
Accepted: 8 January 2024
Published: 11 January 2024

1. Introduction

In numerous real-world applications, instances often pertain to multi-class labels. For example, in text classification, a document may be assigned labels such as "education" and "university". While traditional supervised learning primarily addresses instances belonging to a single-class label, multi-label learning tackles the scenario where instances can be associated with multi-class labels. Over time, multi-label learning has been widely employed in diverse research domains, including text categorization [1–3], automatic annotation for multimedia content [4–6], and image annotation [7,8].

The class-imbalance issue in multi-label classification can manifest in two distinct variants [9,10]. Firstly, there is a significant disparity between the number of positive instances and negative instances for a specific class label. Secondly, in some instances, the number of relevant labels is typically fewer than the number of irrelevant labels. To address these issues, the pairwise loss, which optimizes imbalance-specific evaluation metrics like the area under the ROC curve (AUC) and F measure [11,12], proves more effective than the pointwise loss. Consequently, the application of ranking support vector machine (Rank-SVM) minimizes the pairwise approximate ranking loss, thereby addressing the second aspect of the class-imbalance issue in multi-label classification. This approach effectively mitigates the negative impact of class imbalance. However, apart from the class imbalance problem, there are additional challenges related to noise in multi-label classification. Data corruption and loss in high-dimensional datasets often lead to noise, which can significantly degrade the performance of multi-label classifiers. Consequently, several studies have proposed effective noise reduction methods. For instance, Wu et al. [13] addressed missing labels in image annotation by improving the consistency between predicted and provided labels to enhance the model. Similarly, Cevikalp et al. [14] utilized a ramp loss to mitigate extreme penalties from incorrect-labels and learned a more resilient loss function. However, majority of multi-label classification methods primarily focus on noise within the label space, neglecting the presence of noise in the input feature space.

Manifold regularization matrix factorization is a technique employed to address noise in the input space. Though matrix factorization [15] can effectively solve the problem of image noise, its extensive focus on the global structure of the data while ignoring the local structural features between samples negatively affect its robustness when dealing with complex data. Manifold learning [16] is introduced to adaptively learn local structural features within samples from low rank representations obtained from matrix factorization, ultimately obtaining the most robust low rank representation. This robust low order representation has better robustness when dealing with complex data. However, the features and classifiers selected through manifold regularization matrix factorization are independent of each other, resulting in suboptimal low-rank representation for multi-label classifiers. Additionally, most of these low-rank methods are linear models, limiting their ability to capture the intricate nonlinear relationship between input and output.

To address these challenges, we present a novel multi-label classification model that integrates manifold regularization matrix factorization and a multi-label classifier within a unified framework. We adopt a joint learning approach to simultaneously learn feature selection and classifier parameters. Through joint optimization, optimal features are selected, resulting in a superior joint learning framework that outperforms individual model predictions. This joint framework effectively determines the optimal low-rank representation, leading to improved classification performance in high-noise data scenarios. Furthermore, the proposed unified framework is transformed via kernel function to incorporate a nonlinear multi-label classifier. The objective functions of the linear joint and standard kernel framework are solved using, the augmented Lagrange multiplier (ALM) and alternating direction method of multipliers (ADMM) techniques to obtain efficient values for the variables.

The overall structure of our proposed model is depicted in Figure 1. This research work makes the following key contributions:

- Introduction of a novel approach that combines feature selection and a multi-label SVM classifier within a multi-label classification framework. This is achieved through the utilization of manifold regularized matrix factorization in our proposed method to identify a robust low-rank subspace within the input space. This results in a more resilient SVM classifier, particularly in scenarios with high-noise samples.
- Extending the proposed framework via a kernel learning approach to capture the nonlinear relationship between inputs and outputs. By taking the derivative of the kernel function, we establish a linear relationship between the kernel derivative and the kernel function itself. This leads to an improved solution for the nonlinear problem.
- To effectively address our non-convex learning framework, we employ the augmented Lagrangian multiplier (ALM) and alternating direction method of multipliers (ADMM) techniques. These optimization methods ensure efficient convergence and provide a solution to the optimization problem.

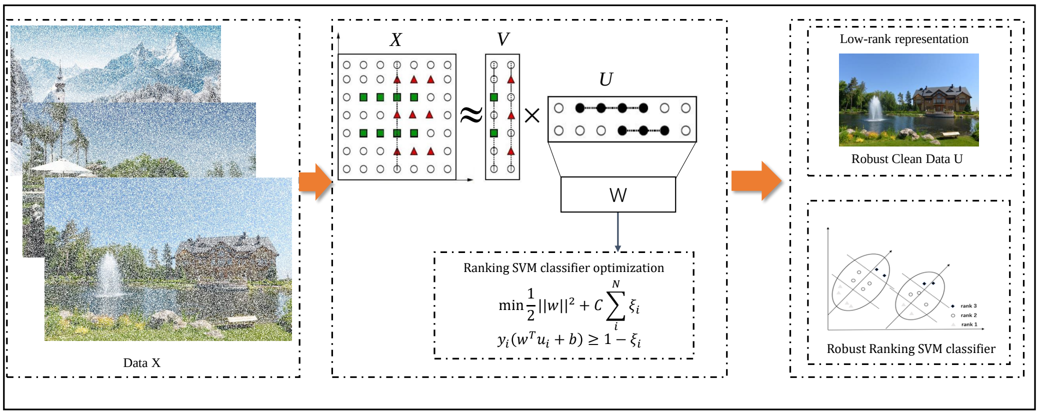
Data X
X
V
U
W
Ranking SVM classifier optimization
$\min \frac{1}{2}||w||^2 + C\sum_{i}^{N} \xi_i$
$y_i(w^T u_i + b) \geq 1 - \xi_i$
Low-rank representation
Robust Clean Data U
rank 3
rank 2
rank 1
Robust Ranking SVM classifier

Figure 1. Illustration of the framework for the proposed robust ranking kernel support vector machine via manifold regularized matrix factorization for multi-label classification.

2. Related Works

2.1. Multi-Label Classification

The current research on machine learning based image classification methods mainly uses machine learning to obtain further information, such as the specific expression of visual image features, image functions, as well as semantic relevance, for the conduct of the research. The purpose of the study is to present a higher level image and adopt a more distinctive way to express the characteristics of the image. However, in actual images, different regions contain different functions. Different functions may correspond to different semantic concepts, or images may have many semantics. This implies that multi-label classification and image recognition are essentially multi-semantic concept classification. Therefore, there are two common ways to solve multiple label classification tasks. One is to divide a multiple classification problem into individual classification problems as in the case of traditional single label classification. For example, BR [4] transforms multiple label classification problems into many independent binary classification problems. Calibrate label sorting (CLR) [17] recasts a multi-label classification task into a label sorting task. Another widely used multi-label classification algorithm is the improved and mature single label classification algorithm [18]. For example, Rank-SVM [19] applies the maximum interval strategy to multi-label classification to minimize sorting loss while maintaining large intervals and using kernel techniques to handle nonlinear situations. ML-KNN [20] originates from the lazy learning technology KNN classifier. However, this method does not consider label correlation in the process of processing multiple semantic tags separately. In existing research on multi-label image classification, it has been found that focusing on the correlation between labels can effectively improve the efficiency of multi-label image classification. Therefore, the ECC combined classifier chain [21] has been improved based on the BR algorithm. The ECC classifier chain links the classifiers that exist in the BR algorithm. The prediction results will be added as part of the attribute vector as a sample. In this way, each binary classifier receives the prediction results of the previous classifier, making full use of the differences between labels. However, this algorithm also has a disadvantage: generating many new tags after calculation. If there is too little training data, allocating sufficient samples for each new tag is impossible, which will likely skew the data. In addition, label combinations that are not included in the training set can significantly reduce the efficiency of classification algorithms.

Although Rank-SVM [19] can minimize the ranking loss, its classification performance is vulnerable to noise, leading to a dramatic decline in its classification performance. This drawback also occurs in other ranking-based methods, such as BP-MLL [22]. Some methods have been proposed to ameliorate this problem, such as calibrated Rank-SVM [23] and Rank-SVMz [24]. However, little work has been conducted to solve this problem by combining feature selection.

2.2. Matrix Factorization

In image classification, machine learning methods commonly include sparse coding, low rank learning, and linear and nonlinear dimensionality reduction techniques. The optimization model in these methods can be considered a constrained matrix factorization problem, aiming to learn the corresponding coefficient matrix or projection matrix through matrix factorization. Our approach seeks to learn a low-dimensional feature representation as the input of multi-label learning, effectively overcoming the negative impact of high-dimensional data and noise on traditional multi-label classification methods. Given a data matrix $X \in \mathbb{R}^{d\times n}$, where d is the feature dimension and n is the number of data samples, we perform the matrix factorization on X. Formally, we derive that

$$\min \left|\left| X - VU^T \right|\right|_F^2, \quad s.t. \quad V^T V = I, \tag{1}$$

where $V \in \mathbb{R}^{d\times r}$ is the basis matrix and $U \in \mathbb{R}^{n\times r}$ is the coefficient matrix, and it can be considered as the new representation of X. The constraint $V^T V = I$ can eliminate redundancy and irrelevant information.

There are many image classification algorithms based on matrix factorization. For example, in [25], sparse coding was applied to natural image classification, and a spatial pyramid matching algorithm using sparse representation was proposed. The classification accuracy on Caltech-101 and Caltech-256 datasets shows that the spatial pyramid matching model based on sparse coding achieves good results. Ref. [26] expands the sparse representation theory and applies it to face recognition. Using training samples as a dictionary, test samples can be represented as linear combinations of base vectors in the dictionary, and further characterized by the degree of sparsity of the combination coefficients. That is, the linear combination coefficients of base vectors consistent with the test sample category tend to be non-zero, while the coefficients of base vectors belonging to other categories are zero. This method obtains sparse coefficients by minimizing the ℓ_1 norm, and uses the minimum reconstruction error to classify face data. However, Ref. [26] does not study dictionaries, and when training and testing samples are simultaneously contaminated with noise, the implementation effect of this method will deteriorate. Ouyang [27] proposed an improved ELM-AE architecture that exploited low-rank matrix decomposition to learn optimal features. In this article, we introduce manifold structure regularization based on matrix factorization. By introducing manifold learning to maintain the manifold relationship between low order samples, we solve the problem of matrix factorization ignoring local structural features between samples, allowing us to obtain the most robust low rank representation. This robust low order representation could be utilized and applied to ultimately obtain better classifiers.

3. Methodology

This section analyzes matrix factorization and ranking support vector machine collaborative learning methods in detail. Firstly, a new multi-label classification framework is constructed based on the matrix factorization method. Then, we propose the kernelization of a linear model.

3.1. Preliminary

In the context of a given matrix, denoted as A, its transpose is represented as A^T. The i-th row and j-th column of A are denoted as a_i and a^j, respectively. The vector ℓ_2-norm of a_i is represented as $||a_i||$ (or $||a_i||_2$). The matrix ℓ_1-norm of A is denoted as $||A||_1$, while the Frobenius norm is represented as $||A||_F$ (or $||A||$). The trace operator for a matrix is denoted as $\mathrm{Tr}(\cdot)$, and the rank of a matrix is represented as $\mathrm{Rank}(\cdot)$. The trace norm (or nuclear norm) of A, denoted as $||A||_*$, is calculated as $\mathrm{Tr}\left(\sqrt{A^T A}\right)$ or as the sum of the i-th largest singular values of A, represented as $\sum_i \sigma_i(A)$.

3.2. Robust Ranking Support Vector Machine

We commence with the fundamental linear Rank-SVM approach [19] method for multi-label classification. For instance, $x \in \mathbb{R}^m$, its real-valued prediction is obtained by $f = xW + b$, where $b = [b_1, b_2, \ldots, b_l] \in \mathbb{R}^l$ is the bias and $W = \left[w^1, w^2, \ldots, w^l\right] \in \mathbb{R}^{m \times l}$ is the parameter matrix. To simplify the formulation, we can absorb b_j into w^j by appending 1 to each instance x as an additional feature. The objective of Rank-SVM is to minimize the ranking loss while maximizing the margin. The ranking learning step can be formulated as follows:

$$\begin{aligned} \min_{U,\,V,\,W,\,b} \ & \frac{1}{2}\sum_{j=1}^{l}\left|\left|w^j\right|\right|^2 + C\sum_{i=1}^{N}\frac{1}{|Y_i^+||Y_i^-|}\sum_{p\in Y_i^+}\sum_{q\in Y_i^-}\xi_{pq}^i \\ s.t. \ & < w^p - w^q, u_i > + b^p - b^q \geq 1 - \xi_{pq}^i, \\ & (p,q) \in Y_i^+ \times Y_i^-, \\ & \xi_{pq}^i \geq 0,\ i = 1, \ldots, n. \end{aligned} \tag{2}$$

where Y_i^+ (or Y_i^-) represents the index set of relevant (or irrelevant) labels associated with the instance x_i. The notation $|\cdot|$ represents the cardinality of a set, and the tradeoff hyper-parameter C is utilized to control the complexity of the model.

However, the performance of the Rank-SVM method is very sensitive to noise data samples and may decline sharply when the noise is high. In this paper, the robustness of feature space is used to solve the impact of noise on classification performance. Given a data matrix $X \in \mathbb{R}^{d \times n}$, where d represents the feature dimension and n denotes the number of data samples. To fully study the impact of noise on data, this paper introduces matrix factorization to solve this problem. We decompose the noisy data X into U and V, where $V \in \mathbb{R}^{d \times r}$ is the basis matrix and $U \in \mathbb{R}^{n \times r}$ is the coefficient matrix, and it can be considered as the new representation of X. Then, training the clean data matrix U into Rank SVM can make our model more robust and resistant to noise.

In addition, since the features selected from the matrix factorization and the multi-label classifier are independent, the low-rank representation is not optimal for the multi-label classification. This paper proposes a new multi-label classification model to improve the classification performance in high noisy data scenarios, seamlessly integrating the matrix factorization and multi-label classifier into a unified framework. Specifically, we use joint learning to combine feature selection and classifier learning, to learn each other's parameters, and finally select the best features through joint optimization, thus realizing a joint learning framework. The robust ranking support vector machine via matrix factorization is as follows:

$$\begin{aligned} \min_{U,\,V,\,E,\,W,\,S,\,b} \ & \frac{\alpha}{2}||E||_{2.1} + \frac{1}{2}\sum_{i,j}||u_i - u_j||_2^2 s_{ij} \\ & + \gamma||U||_* + \theta||S||_* + \frac{1}{2}\sum_{p=1}^{l}||w^p||^2 \\ & + C\sum_{i=1}^{N}\frac{1}{|Y_i^+||Y_i^-|}\sum_{p\in Y_i^+}\sum_{q\in Y_i^-}\xi_{pq}^i \\ s.t. \ & E = X - VU^T,\ \ V^TV = I, \\ & \mathrm{diag}(S) = 0,\ S \geq 0,\ S^T\mathbf{1} = 1, \\ & < w^p - w^q, u_i > +b^p - b^q \geq 1 - \xi_{pq}^i, \\ & (p,q) \in Y_i^+ \times Y_i^-,\ \xi_{pq}^i \geq 0,\ i = 1,\ \ldots,\ n \end{aligned} \tag{3}$$

where S is the similarity matrix, in general, a smaller distance $||u_i - u_j||$ corresponds to a larger weight S_{ij}, defined as $\mathrm{diag}(S) = 0$ and $S \geq 0$, which implies $S_{ii} = 0$ and $S_{ij} \geq 0$. We introduced manifold learning into the proposed framework to maintain manifold relationships between low rank samples to enhance robust low rank representations. With this robust low rank representation, we can find robust low-order features and apply them to obtain better classifiers. $V^T V = I$ can eliminate redundancy and irrelevant information.

3.3. Kernelization

The model described in Equation (4) is a linear multi-label classifier, limiting its effectiveness in capturing nonlinear relationships between the input and output. To address this limitation, we propose the utilization of kernel methods to develop nonlinear multi-label classifiers. The kernel function is then introduced into our proposed framework via the dual formulation of the problem in Equation (4) using the Karush–Kuhn–Tucker theorem. Dual variables are then added to transform the constrained problem into an unconstrained problem which can be optimized using the Lagrangian function as shown below:

$$\begin{aligned}
L(U,\ V,\ E,\ S,\ w,\ \xi,\ \tau,\ \eta) = {} & \frac{\alpha}{2}||E||_{2.1} + \frac{1}{2}\sum_{i,j}||u_i - u_j||_2^2 s_{ij} \\
& + \gamma||U||_* + \theta||S||_* + \frac{1}{2}\sum_{p=1}^{l}||w^p||^2 \\
& + C\sum_{i=1}^{N}\frac{1}{|Y_i^+||Y_i^-|}\sum_{p\in Y_i^+}\sum_{q\in Y_i^-}\xi_{pq}^i \\
& + \frac{u}{2}\left|\left|E - X + VU^T + \frac{M_1}{u}\right|\right|_F^2 \\
& - \sum_{i=1}^{N}\sum_{p\in Y_i^+}\sum_{q\in Y_i^-}\tau_{pq}^i\left(< w^p - w^q, u_i > + b^p - b^q - 1 + \xi_{pq}^i\right) \\
& - \sum_{i=1}^{N}\sum_{p\in Y_i^+}\sum_{q\in Y_i^-}\eta_{pq}^i\xi_{pq}^i
\end{aligned} \tag{4}$$

We now seek a saddle point of the Lagrangian, which would be the minimum for the primal variables $\{w, b, \xi\}$ and the maximum for the dual variable τ. To find the minimum over the primal variables we require,

$$\begin{aligned}
\partial_{b^p} L = 0 \quad & \Longrightarrow \quad \sum_{i=1}^{N}\sum_{j\in Y_i^+}\sum_{q\in Y_i^-} c_{jq}^i \tau_{jq}^i = 0 \\
c_{jq}^i = & \begin{cases} 0 & \text{if } j \neq p \text{ and } q \neq p, \\ +1 & \text{if } j = p, \\ -1 & \text{if } q = p. \end{cases}
\end{aligned} \tag{5}$$

Similarly, for ξ we require

$$\partial_{\xi_{pq}^i} L = 0 \quad \Longrightarrow \quad \frac{C}{|Y_i^+||Y_i^-|} = \tau_{pq}^i + \eta_{pq}^i \tag{6}$$

Similarly, for w we require

$$\partial_{w^p} L = 0 \quad \Longrightarrow \quad w^p = \sum_{i=1}^{N}\left(\sum_{j\in Y_i^+}\sum_{q\in Y_i^-} c_{jq}^i\,\tau_{jq}^i\right)u_i \tag{7}$$

Let $\varphi(\cdot)$ be a feature mapping function that maps x from $\mathbb{R}^d$ to a Hilbert space H. Consequently, the optimization problem described in Equation (4) can be reformulated as follows:

$$\begin{aligned}\min_{U,\ V,\ E,\ W,\ S,\ b}\ \max_{\tau^i_{jq}}\ & \frac{\alpha}{2}||E||_{2.1}+\frac{1}{2}\sum_{i,j}||u_i-u_j||_2^2 s_{ij}\ +\ \gamma||U||_* \\ & +\theta||S||_*-\frac{1}{2}\sum_{p=1}^{l}\sum_{h,i=1}^{N}\beta^h_p\beta^i_p\varphi(u_h)\varphi(u_i)^T \\ & +\sum_{i=1}^{N}\sum_{p\in Y_i^+}\sum_{q\in Y_i^-}\tau^i_{pq} \\ s.t.\quad & E=X-VU^T,\quad V^TV=I \\ & \text{diag}(S)=0,\quad S\geq 0,\quad S^T1=1 \\ & \tau^i_{jq}\in[0,C],\quad \sum_{i=1}^{N}\sum_{j\in Y_i^+}\sum_{q\in Y_i^-}c^i_{jq}\tau^i_{jq}=0, for\ p=1,\ldots,l\end{aligned}\tag{8}$$

where $\beta^i_p=\sum\limits_{j\in Y_i^+}\sum\limits_{q\in Y_i^-}c^i_{jq}\tau^i_{jq}$ and $w^p=\sum\limits_{i=1}^{N}\beta^i_p u_i$

Define $K(u_i,u_j)=\varphi(u_i)\varphi(u_j)^T$ to be the kernel matrix (or Gram matrix) in the RKHS. Consequently, the kernel framework is established as follows

$$\begin{aligned}\min_{U,\ V,\ E,\ W,\ S,\ b}\ \max_{\tau^i_{jq}}\ & \frac{\alpha}{2}||E||_{2.1}+\frac{1}{2}\sum_{i,j}||u_i-u_j||_2^2 s_{ij}\ +\ \gamma||U||_* \\ & +\theta||S||_*-\frac{1}{2}\sum_{p=1}^{l}\sum_{h,i=1}^{N}\beta^h_p\beta^i_pK(u_i,u_j) \\ & +\sum_{i=1}^{N}\sum_{p\in Y_i^+}\sum_{q\in Y_i^-}\tau^i_{pq} \\ s.t.\quad & E=X-VU^T,\quad V^TV=I \\ & \text{diag}(S)=0,\quad S\geq 0,\quad S^T1=1 \\ & \tau^i_{jq}\in[0,C],\quad \sum_{i=1}^{N}\sum_{j\in Y_i^+}\sum_{q\in Y_i^-}c^i_{jq}\tau^i_{jq}=0,\quad \text{for } p=1,\ldots,l\end{aligned}\tag{9}$$

3.4. Optimization

Since our model involves the $\ell_{2.1}$-norm, the problem's orthogonal and nonnegative constraints are still challenging. Therefore, this article proposes a new and effective optimization method based on ALM to solve the problem. We aim to introduce auxiliary variables to separate constraints and keep their equivalence during optimization. Specifically, we introduce two auxiliary variables $H=S$ and $Z=U$ and transform the objective function into the following form

$$\begin{aligned}& L(U,\ V,\ E,\ S,\ Z,\ H,\ w,\ \xi,\ \tau,\ \eta\ M_1,\ M_2,\ M_3) \\ & =\frac{\alpha}{2}||E||_{2.1}+\frac{1}{2}tr\left(U^TL_sU\right)+\gamma||Z||_* \\ & +\frac{u}{2}\left|\left|E-X+VU^T+\frac{M_1}{u}\right|\right|_F^2+\frac{u}{2}\left|\left|Z-U+\frac{M_2}{u}\right|\right|_F^2 \\ & +\frac{u}{2}\left|\left|H-S+\frac{M_3}{u}\right|\right|_F^2+\theta||H||_*\end{aligned}\tag{10}$$

$$
\begin{aligned}
&- \frac{1}{2}\sum_{p=1}^{l} \sum_{h,i=1}^{N} \beta_p^h \beta_p^i K(u_i, u_j) + \sum_{i=1}^{N} \sum_{p \in Y_i^+} \sum_{q \in Y_i^-} \tau_{pq}^i \\
s.t.\quad & V^T V = I, \quad \mathrm{diag}(S) = 0, \quad S \geq 0, \quad S^T \mathbf{1} = 1 \\
& \tau_{jq}^i \in [0, C], \quad \sum_{i=1}^{N} \sum_{j \in Y_i^+} \sum_{q \in Y_i^-} c_{jq}^i \tau_{jq}^i = 0, \quad \text{for } p = 1, \ldots, l
\end{aligned}
$$

where u is the regularization parameter that determines the penalty for infeasibility, $M_1 \in \mathbb{R}^{d \times n}$, $M_2 \in \mathbb{R}^{n \times r}$ and $M_3 \in \mathbb{R}^{n \times n}$ are the Lagrangian multipliers that penalize the gap between the target and the auxiliary variable.

With the transformation, we can adopt alternative optimization to iteratively solve the problem. Specifically, we optimize the objective function with respect to one variable while fixing the remaining variables. The iteration steps are detailed as follows.

(1) Update U: By fixing the other variables, the optimization formula for U becomes

$$
\begin{aligned}
\min_{U} \; & \frac{1}{2}\sum_{i,j} ||u_i - u_j||_2^2 S_{ij} + \frac{u}{2}\left|\left|E - X + VU^T + \frac{M_1}{u}\right|\right|_F^2 \\
& + \frac{u}{2}\left|\left|Z - U + \frac{M_2}{u}\right|\right|_F^2 - \frac{1}{2}\sum_{p=1}^{l} \sum_{h,i=1}^{N} \beta_p^h \beta_p^i K(u_i, u_j)
\end{aligned}
\tag{11}
$$

According to Karush–Kuhn–Tucker (KKT) condition and $K(U_i, U_j) = \exp(-\gamma||u_i - u_j||^2)$, it can be verified that the optimal solution should be

$$
\begin{aligned}
u_i = \mathrm{sylvester}\Bigg(& \sum_{j=1}^{N} S_{ij} + u + \gamma \sum_{j=1}^{N} K(U_i, U_j) A_{ij}, \, UV^T V, \\
& \sum_{j=1}^{N} u_j S_{ij} - U\left(E_{i:}^T - X_{i:}^T + \frac{M_{1_{\mathrm{i:}}}^T}{U} \right) V \\
& + U\left(Z_{i:} + \frac{M_{2_{i:}}}{U} \right) + \gamma \sum_{j=1}^{N} K(U_i, U_j)(U_j) A_{ij} \Bigg)
\end{aligned}
\tag{12}
$$

(2) Update V: By fixing the other variables, the optimization formula for V becomes

$$
\min_{V} \; \frac{u}{2}\left|\left|E - X + VU^T + \frac{M_1}{u}\right|\right|_F^2 \tag{13}
$$

Considering $V^T V = I$, the above formula can be rewritten as

$$
\min_{V} \; \left|\left|VU^T - \left(X - E - \frac{M_1}{U}\right)\right|\right|_{\mathbf{F}}^2 \tag{14}
$$

This problem is commonly referred to as the orthogonal procrustes problem, for which the global optimal solution can be obtained through the singular value decomposition of $U^T\left(X - E - \frac{M_1}{u}\right)^T$. To be more specific, given

$$
[V_1 \; S_1 \; U_1] = \mathrm{SVD}\left(U^T \left(X - E - \frac{M_1}{U} \right)^T \right) \tag{15}
$$

The following formula can update V

$$
V = U_1 {V_1}^T \tag{16}
$$

(3) Update E: By fixing the other variables, the optimization formula for E becomes

$$\min_{E} \frac{\alpha}{2}||E||_{2.1} + \frac{U}{2}\left|\left|E - \left(X - VU^T - \frac{M_1}{U}\right)\right|\right|_{\mathbf{F}}^2 \tag{17}$$

Let $T = \left(X - VU^T - \frac{M_1}{u}\right)$, $\tau = \frac{\alpha}{2}$, and we further have

$$\min_{E} \tau||E||_{2.1} + \frac{1}{2}||E - T||_{\mathbf{F}}^2 \tag{18}$$

To solve the above equation, we introduce the following lemma, also presented in [28], with detailed proof.

Lemma 1. *Given a matrix $W = \{w_1, w_2, \ldots, w_n\} \in \mathbb{R}^{m\times n}$ and a positive scalar λ, then X^* is the optimal solution of*

$$\min_{E} \lambda||X||_{2.1} + \frac{1}{2}||X - W||_{\mathbf{F}}^2 \tag{19}$$

and the column of X^*

$$X_i^* = \begin{cases} \frac{\|W_i\| - \lambda}{\|W_i\|} W_i & \text{, if } \lambda < \|W_i\| \\ 0 & , \text{ otherwise} \end{cases} \tag{20}$$

According to Lemma 1, the solution of the above problem is

$$E(:, i) = \begin{cases} \frac{\|T_i\| - \tau}{\|T_{\mathrm{i}}\|} T_i & \text{, if } \tau < \|T_i\| \\ 0 & , \text{ otherwise} \end{cases} \tag{21}$$

(4) Update Z: By fixing the other variables, the optimization formula for Z becomes

$$\min_{Z} \gamma||Z||_* + \frac{u}{2}\left|\left|Z - U + \frac{M_2}{u}\right|\right|_F^2 \tag{22}$$

We first introduce the soft thresholding (shrinkage) operator

$$S_\varepsilon[q] = \max(|q| - \varepsilon, 0)\mathrm{sgn}(q) \tag{23}$$

where $\mathrm{sgn}(q)$ represents the *sign* function. By applying the shrinkage operator element-wise to the singular values of $U - \frac{M_2}{u}$, the optimal update for Z can be expressed as follows:

$$Z = U_1 S_{\frac{\gamma}{u}}[S_1]V_1^T \tag{24}$$

in which U_1, S_1 and V_1 are the SVD factorization

$$\left[U_1, S_1, V_1^T\right] = \mathrm{SVD}\left(U - \frac{M_2}{u}\right) \tag{25}$$

(5) Update S: By fixing the other variables, the optimization formula for S can be derived as

$$\begin{gathered} \min_{S} \frac{1}{2}\sum_{i,j} ||U_i - U_j||_2^2 S_{ij} + \frac{u}{2}\left|\left|H - S + \frac{M_3}{u}\right|\right|_F^2 \\ s.t. \quad \mathrm{diag}(S) = 0, S \geq 0, S^T\mathbf{1} = 1 \end{gathered} \tag{26}$$

where $\|U_i - U_j\|_2^2 = G_{ij}$, then

$$\min_{S} \frac{1}{2}\mathrm{Tr}\left(G^T S\right) + \frac{\mu}{2}\left\|S - \left(H + \frac{M_3}{\mu}\right)\right\|_F^2 \tag{27}$$

We denote $N = H - \frac{M_3}{\mu}$

$$\begin{aligned} &\min_{S_i} \|S_i - (N_i - G_i/2\mu)\|_2^2 \\ s.t. \quad &\operatorname{diag}(S) = 0, S \geq 0, S^T\mathbf{1} = 1 \end{aligned} \tag{28}$$

Considering the constraints

$$\begin{aligned} \min_{S_i} \sum_s \|S_i - (N_i - G_i/2\mu)\|_2^2 - \eta\left(1^T S_i - 1\right) - \zeta^T S_i \\ s.t. \ \operatorname{diag}(S) = 0 \end{aligned} \tag{29}$$

Taking the derivative with respect to S_i and setting it to zero

$$S_i - (N_i - G_i/2\mu) - \eta 1 - \zeta = 0 \tag{30}$$

The j entry of S_i is shown below

$$S_{ij} - \left(N_{ij} - G_{ij}/2^{-}\right) - \eta - \zeta_i = 0 \tag{31}$$

According to KKT conditions

$$S_{ij} = \left(\left(N_{ij} - G_{ij}/2\mu\right) + \eta\right)_+ \tag{32}$$

(6) Update H: By fixing the other variables, the optimization formula for H becomes

$$\min_H \ \theta||H||_* + \frac{u}{2}\left|\left|H - S + \frac{M_3}{u}\right|\right|_{\mathbf{F}}^2 \tag{33}$$

To begin, we introduce the soft thresholding (shrinkage) operator

$$S_\varepsilon[q] = \max(|q| - \varepsilon, 0)\operatorname{sgn}(q) \tag{34}$$

where $\operatorname{sgn}(q)$ represents the *sign* function. By applying the shrinkage operator to the singular values of $S - \frac{M_3}{u}$ element-wisely, the optimal updation of H is given by

$$H = U_1 S_{\frac{\theta}{u}}[S_1]V_1^T \tag{35}$$

in which U_1, S_1 and V_1 are the SVD factorization

$$\left[U_1,\ S_1,\ V_1^T\right] = \operatorname{SVD}\left(S - \frac{M_3}{u}\right) \tag{36}$$

(7) Update τ: By fixing the other variables, the optimization formula for τ becomes

$$\begin{aligned} &\max_{\tau_{jq}^i} - \frac{1}{2}\sum_{p=1}^{l}\sum_{h,i=1}^{N} \beta_p^h \beta_p^i K(u_i, u_j) + \sum_{i=1}^{N}\sum_{p\in Y_i^+}\sum_{q\in Y_i^-} \tau_{pq}^i \\ &s.t. \ \ \tau_{jq}^i \in [0, C], \sum_{i=1}^{N}\sum_{j\in Y_i^+}\sum_{q\in Y_i^-} c_{jq}^i \tau_{jq}^i = 0, \text{for } p = 1, \ldots, l \end{aligned} \tag{37}$$

Then, we can obtain the optimal solution of Equation (37) through the general Quadratic programming solution method or SMO algorithm.

(8) Update u, M_1, M_2 and M_3: Finally, we need to update the ALM parameters

$$M_1 = M_1 + u\left(E - X + VU^T\right) \tag{38}$$

$$M_2 = M_2 + u(Z - U) \tag{39}$$

$$M_2 = M_2 + u(Z - U) \tag{40}$$

$$u = \rho u \tag{41}$$

where the parameter $\rho > 1$ is the learning rate that controls the convergence speed. The flowchart of the proposed approach is presented in Algorithm 1.

Algorithm 1 The proposed approach.

```
Input: Training set $\{X_1, \ldots, X_n\} \in \mathbb{R}^{d\times n}$, parameter $\alpha$, $\gamma$, $\theta$
Initialize: $M_1 = M_2 = M_3 = 0, U = 0, V = 0, E = 0, Z = 0, S = 0, H = 0$
while not converged do
    Fix others and update $U$ by Equation (12)
    Fix others and update $V$ by Equation (16)
    Fix others and update $E$ by Equation (21)
    Fix others and update $Z$ by Equation (24)
    Fix others and update $S$ by Equation (32)
    Fix others and update $H$ by Equation (35)
    Fix others and update $\tau$ by solving Equation (37)
end while
Output: The correlation matrices $U$, $V$ and $\tau$.
```

4. Experiments

In this section, we verified the effectiveness of our proposed model by conducting a series of comparative experiments on six widely used multi-label benchmark data sets with five states of the art multi-label classification methods. First, a brief description of the dataset as well as the experimental evaluation metrics are provided. Five advanced multi-label classification methods are selected and discussed. Comparative experiments were conducted on standard dataset with and without noise. The classification performance of the proposed algorithm as well as the comparative methods were analysed, evaluated and reported accordingly.

4.1. Datasets and Experimental Setting

Datasets description: To evaluate the performance of our proposed method, we selected several multi-label data sets in different scenarios in the real world for experiments. Table 1 summarizes the details of the selected multi-label data sets comprising multiple data containing multi-variable characteristics and labels. Therefore, they are a group of rather complex data, resulting in increased computational complexity in the learning process.

- **Image** contains 2000 samples from the image domain, and these samples have 5 labels.
- **Scene** contains 2407 samples from the image domain, and these samples have 6 labels.
- **Yeast** contains 2417 samples from the biology domain, and these samples have 14 labels.
- **Cal500** contains 502 samples from the music domain, and these samples have 174 labels.
- **Rcv1s1** contains 6000 samples from the text domain, and these samples have 101 labels.
- **Core116k4** contains 13,837 samples from the image domain, and these samples have 162 labels.

Evaluation metrics: In evaluating the performance of multi-label classification learning methods, the correlation between multiple labels for each instance must be considered. This makes the evaluation of muti-label classification cumbersome compared to traditional single-label classification. The evaluation methods used in current classification learning tasks can be categorized into sample-based and label-based measurements. The label-based indicators are designed to calculate the metrics for each individual label and then aggregate them to obtain the overall label indicators using either macro or micro averaging techniques. In addition, in multi-label classification learning, the measure of evaluation by each instance is generally used. The multi-label classification algorithm mainly calculates the inconsistent ratio between the predicted label and the actual label of each instance. Then the evaluation is performed with metrics such as Hamming loss, accuracy, F1 measure, ranking loss, etc.

Table 1. Statistics of the experimental datasets. ("Cardinality" indicates the average number of labels per example. "Density" normalizes the "Cardinality" by the number of possible labels. "URL" indicates the source URL of the dataset).

Datasets	Examples	Features	Labels	Cardinality	Density	Domain	URL
Image	2000	294	5	1.240	0.248	image	URL 2
Scene	2407	294	6	1.074	0.179	image	URL 1
Yeast	2417	103	14	4.237	0.303	biology	URL 1
Cal500	502	68	174	26.044	0.150	Music	URL 1
Rcv1s1	6000	944	101	2.880	0.029	Text	URL 1
Core116k4	13,837	500	162	2.867 ± 0.033	0.018 ± 0.001	image	URL 1

URL 1: http://mulan.sourceforge.net/datasets-mlc.html, accessed on 8 July 2023. URL 2: http://palm.seu.edu.cn/zhangml, accessed on 8 July 2023.

In this paper, we used the following five evaluation metrics for the multi-label learning classification. Given a test set $D = \{(X_i, Y_i), 1 \leq i \leq m\}$, where Y represents the label datasets of dimension q and X represents the instance datasets of dimension p.

(1) Hamming loss quantifies the proportion of example-label pairs that are misclassified.

$$\text{Hammingloss} = \frac{1}{m}\sum_{i=1}^{m}\frac{|h(X_i)\Delta Y_i|}{q} \tag{42}$$

where Δ denotes the symmetric difference between two sets. $h(X_i)$ represents the trained predictive function that generates a set of predictive labels based on the value of X_i. Therefore, a lower Hamming loss value indicates a higher degree of similarity between the predicted and actual values, signifying improved classifier performance.

(2) Subset accuracy measures the fraction that the predicted label subset and the ground-truth label subset are the same.

$$\text{Subset Accuracy} = \frac{1}{m}\sum_{i=1}^{m}[h(X_i) = Y_i] \tag{43}$$

(3) Example-F1 is the average F1 measure that is the harmonic mean of recall and precision over each instance.

$$\text{Example-F1} = \frac{1}{m}\sum_{i=1}^{m}\frac{2|P_i^+ \cap Y_i^+|}{|P_i^+|+|Y_i^+|} \tag{44}$$

where Y_i^+(or Y_i^-) represents the index set of the ground-truth relevant (or irrelevant) labels associated with X_i, and P_i^+(or P_i^-) denotes the index set of the predicted relevant (or irrelevant) labels associated with X_i.

(4) Micro-F1 is an indicator used to measure the accuracy of multivariate classification learning. It first averages the element values corresponding to each confusion matrix, yields the TP, FP, TN, and FN, and then calculates the micro-precision and micro-recall.

$$\text{Micro-F1} = B\left(\sum_{i=1}^{q} tp_i, \sum_{i=1}^{q} fp_i, \sum_{i=1}^{q} tn_i, \sum_{i=1}^{q} fn_i\right) \tag{45}$$

where tp_i and fp_i represent each sample's true positive and false positive, respectively, and fn_i and tn_i represent the false negative and true negative.

(5) Ranking loss measures the average fraction of the label pairs that an irrelevant label ranks higher than a relevant label over each instance.

$$\text{Ranking Loss} = \frac{1}{m}\sum_{i=1}^{m}\frac{|SetR_i|}{|Y_i^+||Y_i^-|} \tag{46}$$

where $SetR_i = \{(p,q)|f_p(x_i) \leq f_q(x_i), (p,q) \in Y_i^+ \times Y_i^-\}$.

Selecting the most suitable comparative algorithm based on the specific field of research and available resources can help demonstrate the effectiveness and advantages of the method proposed in this paper. We considered the most representative RANK-SVM algorithm from traditional multi label classification algorithms, and ML-KNN and CPNL algorithms from SVM's unique multi label extensions. We also looked at MLR and RBRL algorithms from the latest multi label methods. The evaluation performance of our proposed algorithm is compared with several excellent multi-label classification methods listed below.

- **Rank-SVM** [19] is an adaption of the maximum margin strategy for MLC.
- **ML-KNN** [20] is a multi-label delayed learning method based on the traditional K-nearest neighbor (KNN) algorithm.
- **CPNL** [29] is a recent method proposed as a cost-sensitive loss function to address the issue of class imbalance and leverages correlations between negative and positive labels to improve performance.
- **MLR** [30] converts MLC task into the pairwise label ranking problem.
- **RBRL** [31] (robust low-rank learning) is a technique that integrates Rank-SVM and binary relevance. Additionally, the thresholding step is incorporated into the ranking learning component of Rank-SVM via binary relevance. This enables model training to occur in a single unified step, rather than sequentially.

4.2. Results under Original Data

In this section, the experimental results of our proposed method and the above-mentioned multi-label classification algorithms are analyzed and discussed in detail. We evaluated the performance of our proposed method and the other comparative multi-label classification algorithms on different data sets using Example-F1, Micro-F1 and Hamming loss evaluation metrics. The specific results of this experimental evaluation are summarized in Tables 2–4. In terms of the evaluation index (↑), a higher value represents better performance of the model. Conversely, (↓) indicates that a lower value corresponds to better model performance. In addition, Table 5 summarizes the average ranking of these comparative methods by each indicator on all data sets, while Figure 2 intuitively illustrates the overall average ranking of these comparative methods on all indicators.

Table 2. Comparative results of competitive algorithms on the datasets using the Example-F1 metric (↑).

Example-F1	Rank-SVM	ML-KNN	CPNL	MLR	RBRL	Proposed
Image	0.3401	0.4278	0.3234	0.6423	0.6455	**0.6573**
Scene	0.5487	0.6473	0.6276	0.7379	0.7412	**0.7444**
Yeast	0.6034	0.6042	0.6277	0.6228	0.6421	**0.6562**
Cal500	0.3442	0.2758	0.2841	0.3882	0.3881	**0.3912**
Rcv1s1	0.3145	0.3056	0.3427	0.3459	0.3523	**0.3607**
Core116k4	0.0712	0.0823	0.0812	0.0924	0.0944	**0.0976**

Table 3. Comparative results of competitive algorithms on the datasets using the Micro-F1 metric (↑).

Micro-F1	Rank-SVM	ML-KNN	CPNL	MLR	RBRL	Proposed
Image	0.4267	0.6015	0.3702	0.6457	0.6477	**0.6572**
Scene	0.6203	0.5961	0.6177	0.7298	0.7354	**0.7501**
Yeast	0.6123	0.5676	0.6425	0.6541	0.6538	**0.6565**
Cal500	0.3369	0.3064	0.2502	0.3846	0.3869	**0.3962**
Rcv1s1	0.3201	0.3342	0.3014	0.3792	0.3811	**0.3824**
Core116k4	0.1258	0.1023	0.0964	0.1396	0.1349	**0.1425**

Table 4. Comparative results of competitive algorithms on the datasets using the Hamming loss metric (↓).

Hamming loss	Rank-SVM	ML-KNN	CPNL	MLR	RBRL	Proposed
Image	0.2869	0.3174	0.3402	0.1768	0.1712	**0.1695**
Scene	0.1427	0.1203	0.1557	0.1143	0.1117	**0.1103**
Yeast	0.2314	0.3104	0.2253	0.2027	0.2019	**0.2006**
Cal500	0.1582	0.1732	0.1794	0.1763	0.1642	**0.1578**
Rcv1s1	0.0413	0.0326	0.0415	0.0397	0.0326	**0.0304**
Core116k4	0.0214	0.0212	0.0463	0.0218	0.0245	**0.0201**

Table 5. Average ranks of the comparative methods on all datasets in terms of each evaluation metric.

Average Rank	Rank-SVM	ML-KNN	CPNL	MLR	RBRL	Proposed
Example-F1	5.50	4.83	4.66	2.83	2.16	**1.00**
Micro-F1	4.50	5.00	5.50	2.66	2.33	**1.00**
Hamming loss	4.66	4.66	5.66	2.83	2.16	**1.00**
Overall	4.88	4.83	5.27	2.77	2.21	**1.00**

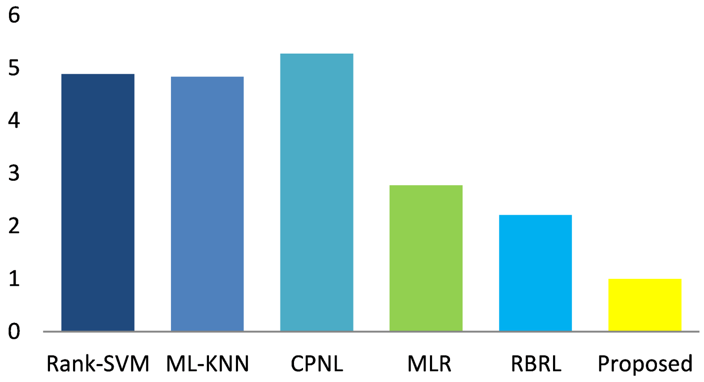

Figure 2. Overall average ranks of the competitive methods on all evaluation metrics.

Table 2 presents the evaluation performance of each comparative algorithm on the dataset using Example-F1 evaluation metric. From the results in the table, it can be observed that, compared to the other multi-label classification methods, our proposed method demonstrated superior performance with respect to the Example-F1 evaluation index on the six multi-label datasets. In particular, the evaluation performance of the proposed algorithm on the Image, Scene, Yeast, Cal500, Rcv1s1 and Core116k4 datasets is improved by 11.4%, 9.41%, 10.34%, 1.3% and 0.73%, respectively, compared with the other comparative algorithms.

Similarly, Table 3 captures the performance of our proposed method and that of the comparative algorithms using the Micro-F1 evaluation metric. In images, scenes, and yeast datasets, the algorithm proposed in this paper obtained a better performance, whilst RBRL and MLR algorithms also demonstrated relatively good performance. This can be attributed to the fact that, RBRL algorithm combines sorting support vector machines and binary correlation with robust low-order learning, whereas the MLR algorithm is systematically studied from two complementary perspectives: the consistency of learning algorithm and the generalized error bound. These two make the classifier more robust when dealing with complex datasets. Our proposed algorithm recorded performance values which are 1.15%, 2.03% and 0.24% higher than that of MLR algorithm, and 0.95%, 1.47% and 0.27% higher than the RBRL algorithm, respectively. The proposed algorithm's ability to combine feature selection with a multi label SVM classifier culminates in the generation of a more robust SVM classifiers under complex samples. In a nutshell, with respect to the Micro-F1 evaluation metric, our method demonstrated a significant superior experimental results in high-dimensional feature space.

The performance analysis of our proposed method and the other comparative algorithms on the selected experimental dataset with respect to the Hamming loss metric is recorded in Table 4. Hamming loss is used to examine the misclassification of samples on a single marker, where relevant markers do not appear in the predicted marker set or irrelevant markers appear in the predicted marker set. Generally, our proposed method significantly outperformed the other comparative algorithms in almost all the dataset used for the experiment. However, the improvement was not significant on the Core116k4 dataset, which may be largely attributed to the large sample size of the dataset.

As shown in Table 5 and Figure 2, the algorithm proposed in this paper achieved better performance than other methods in terms of overall indicators. In summary, compared to several state-of-the-art MLC methods, our proposed algorithm achieved superior performance.

Furthermore, the experimental results on the performance of the RBF and polynomial kernels on the Image, Scene, Yeast, Cal500, Rcv1s1, and Core116k4 datasets with respect to Example-F1, Micro-F1 and Hamming loss evaluation metrics are illustrated in Figure 3. It can be seen from Figure 3 that, the Gaussian kernel achieved better evaluation metric results for the Image, Scene, Yeast and Cal500 datasets while the polynomial kernel performed better in the case of Rcv1s1 and Core116k4 datasets. However, the Gaussian kernel function, which is considered a universal kernel, also performed well on Rcv1s1 and Core116k4 datasets. This implies that, in instances where the number of features is small and the number of samples is normal, Gaussian kernel function is suitably selected. On the other hand, the polynomial kernel may be a better choice in situations where the number of features and samples are large.

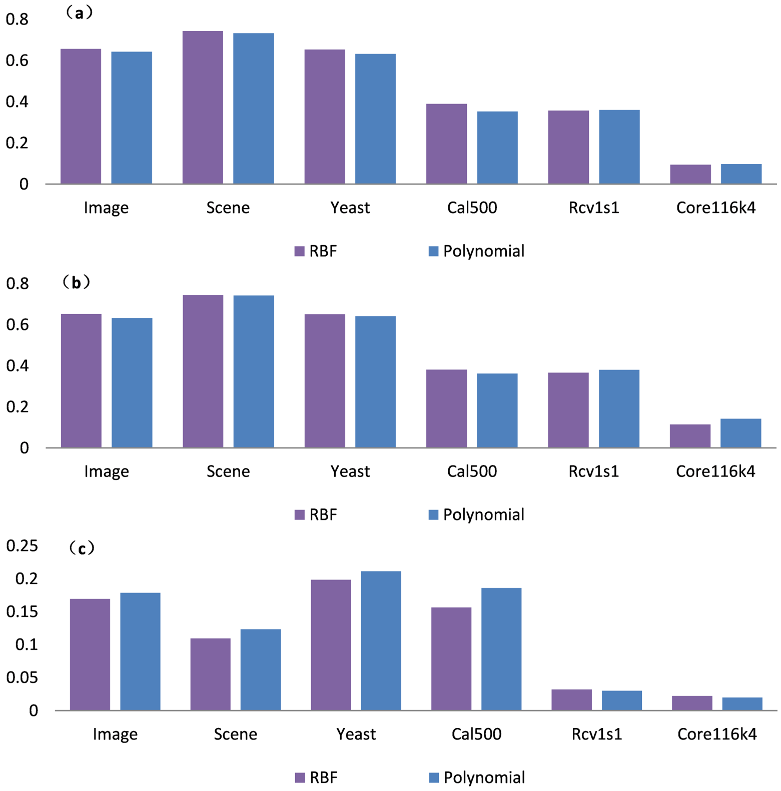

Figure 3. Performance of each kernel on the datasets. (**a**) Example-F1, (**b**) Micro-F1, (**c**) Hamming loss.

4.3. Results under Different Noises

To evaluate the effectiveness of our proposed approach on noisy datasets, we introduced manifold regularization matrix factorization. This enhancement enables our method to handle multi-label classification tasks even when the dataset is affected by noise interference.

Figure 4 presents the performance of our proposed method as well as the other comparative multi-label classification methods on salt and pepper noise induced image, scene, yeast, Rcv1s1 and Core116k4 datasets with respect to Example-F1, Micro-F1 and Hamming loss evaluation metrics. In the noise induced image and scene datasets, the proposed algorithm obtained a better performance compared to the relatively good performance achieved by RBRL and MLR algorithms. This may be attributed to the fact that, RBRL algorithm combines sorting support vector machines and binary correlation with robust low-order learning, which can achieve the effect of resisting noise to a certain extent. The MLR algorithm is systematically studied from two complementary perspectives: the consistency of learning algorithm and the generalized error bound. It can also solve the error problem caused by data corruption to a certain extent. CPNL, ML-KNN and Rank-SVM algorithms showed poor performance in image datasets with noise because they focus too much on label space and ignore the impact of noise on classifiers. In the case of the Yeast dataset without the noise, the difference in performance between the proposed method and other algorithms is not significant. This is because, these methods can still learn prominent or relevant features of the Yeast image. However, as the noise level increases, the effectiveness of our proposed method becomes increasingly obvious. Likewise, with low noise level in the Rcv1s1 and Core116k4 datasets, the evaluation performance of the proposed method compared to RBRL and MLR algorithms is not obvious, but in the case of at least 20% noise pollution, the proposed method demonstrates superior performance than other algorithms with respect to the three evaluation indicators of Example-F1, Micro-F1 and Hamming loss. This is due to the introduction of the manifold regularization matrix factorization in the proposed method, capable of finding robust low-rank subspaces in the input space to demonstrate strong robustness on high-noise datasets.

Figure 5 shows the corresponding changes in the Gaussian noise on the Image, Scene, Yeast, Rcv1s1, and Core116k4 datasets with respect to the performance of Example-F1, Micro-F1, and Hamming loss evaluation metrics. In the case of the Image dataset, a trend similar to salt and pepper noise scenario was observed. Thus, the proposed algorithm, RBRL, and MLR, performed appreciably in noisy environments because all three methods consider the error problem caused by data corruption. Furthermore, due to the introduction of matrix decomposition in the algorithm proposed in this article, robust low-rank subspaces are considered in the input space. As a result, the proposed algorithm is less susceptible to Gaussian noise, hence the superior results. In the Scene and Yeast datasets, the CPNL algorithm also achieved good results, which may be attributed to its extension of BR to solve the problems of category imbalance and label correlation, making the model more robust. However, under different noise levels, our proposed algorithm still have the lowest reduction rate. The algorithm significantly decreases with increasing noise density in the Rcv1s1 and Corel16k7 datasets. However, our method achieved better results than the other algorithms with respect to the evaluation metrics. This is indicative of the fact that, our proposed method has been tested on datasets with Gaussian noise and exhibits good robustness. The experimental results further demonstrate that this method can effectively improve the performance of multi label classifiers.

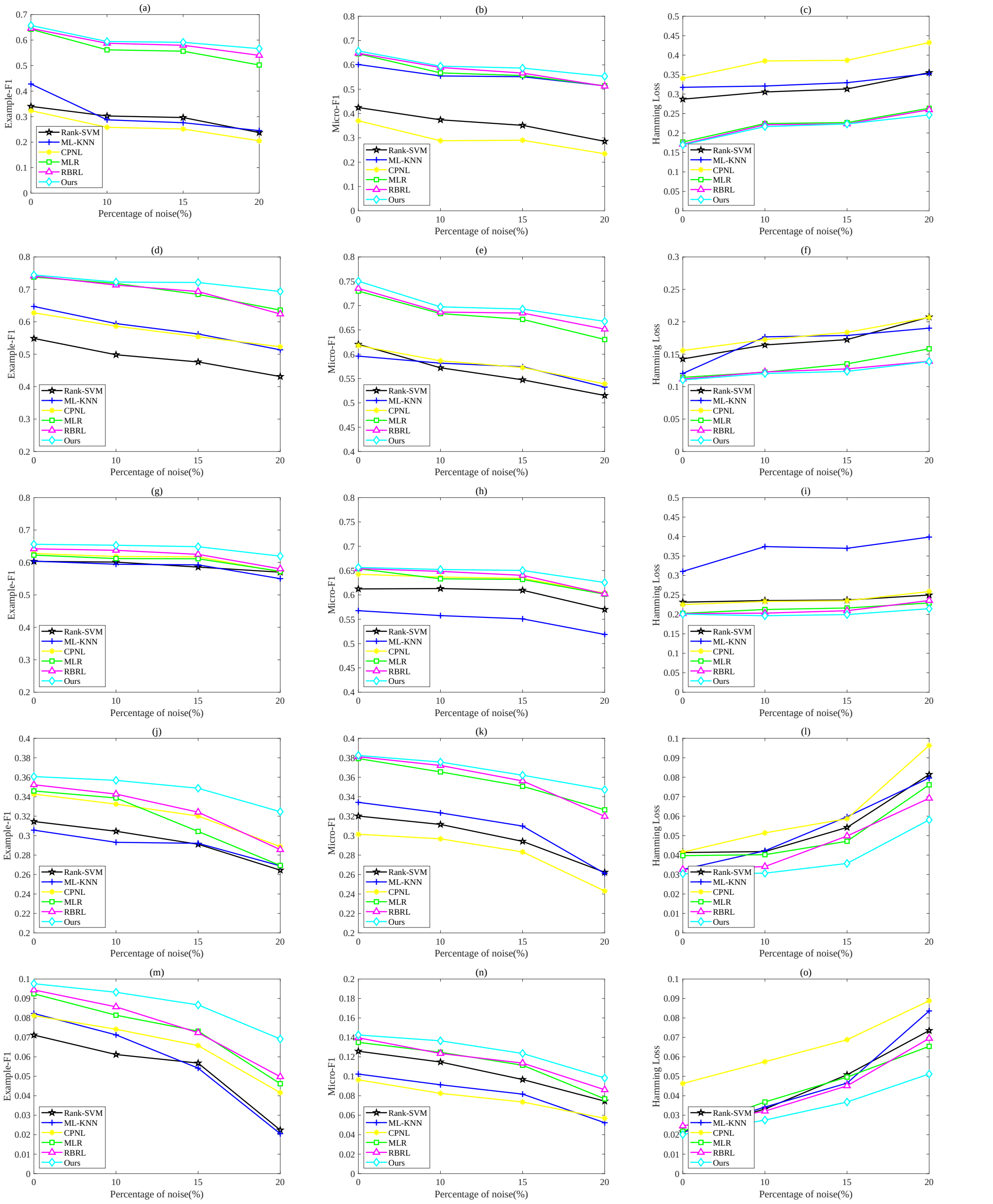

Figure 4. Mean Example-F1, Micro-F1 and Hamming loss of different methods corrupted by varying levels of random "salt & pepper" noise on (**a**–**c**) Image, (**d**–**f**) Scene, (**g**–**i**) Yeast, (**j**–**l**) Rcv1s1, (**m**–**o**) Core116k4 datasets.

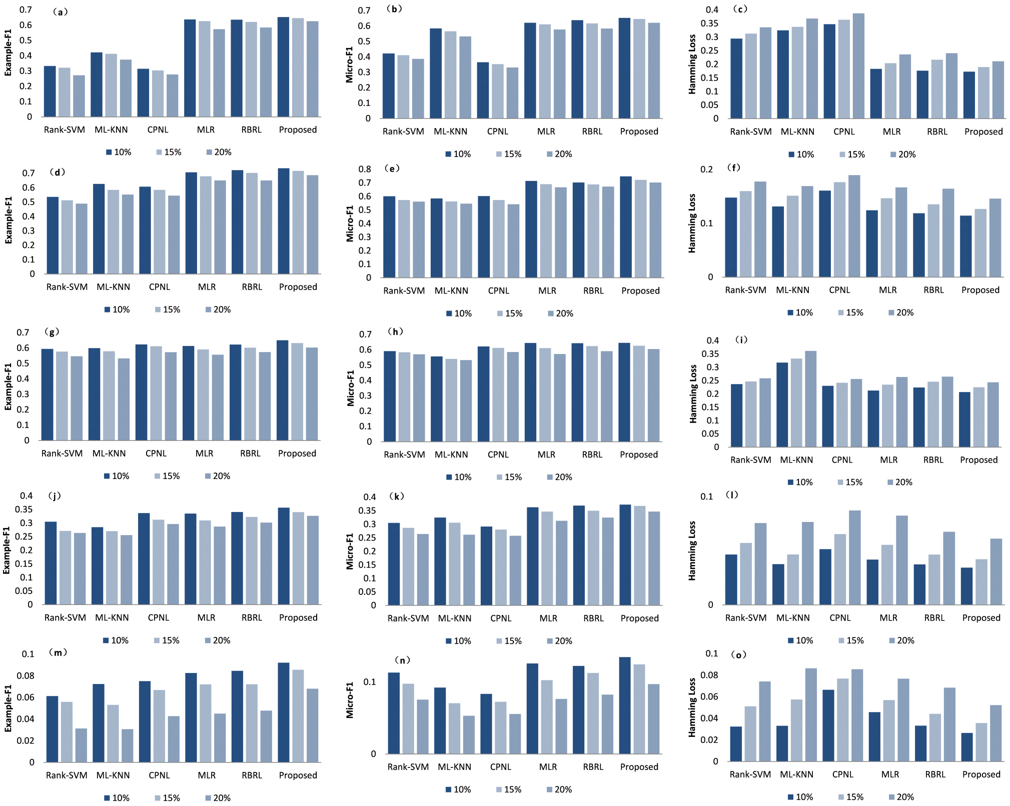

Figure 5. Mean Example-F1, Micro-F1 and Hamming loss of different methods corrupted by varying levels of random "Gaussian" noise on (**a**–**c**) Image, (**d**–**f**) Scene, (**g**–**i**) Yeast, (**j**–**l**) Rcv1s1, (**m**–**o**) Core116k4 datasets.

4.4. Parameter Sensitivity and Convergence

In this section, we conducted parameter sensitivity analysis on the proposed method. We analyze the sensitivity of this method to the regularization parameters α, γ and θ in the loss function Equation (10). Figure 6 shows the impact of these regularization parameters on the Example-F1 indicator of our proposed method on the Scene dataset. In Figure 6a, we determine $\alpha = 1$ and change γ and θ within a similar range $\left[10^{-2}, 10^{4}\right]$. When $\theta = 1$, the value of γ does not affect the experimental effect. In Figure 6b, we determine that $\gamma = 1$. When $\alpha = 10$ and $\theta = 1$, the experimental results are good. In Figure 6c, we determine $\theta = 1$. When $\alpha = 1$ and $\gamma = 10$, the results are better. In Figure 6d–f, When $\alpha = 0.1$, $\gamma = 10$ and $\theta = 1$, The Micro-F1 indicator yields better results. In Figure 6g–i, When $\alpha = 100$, $\gamma = 100$ and $\theta = 100$, The Hamming-Loss indicator yields better results. From the figure, it can be seen that when we keep the values of these three parameters within a specific range, the final performance of the method on Example-F1, Micro-F1 and Hamming-Loss is not sensitive to the selection of parameters. It can be seen from Figure 6 that the proposed method is very sensitive to hyperparameters. In addition, the performance is best at some intermediate values of α, γ and θ.

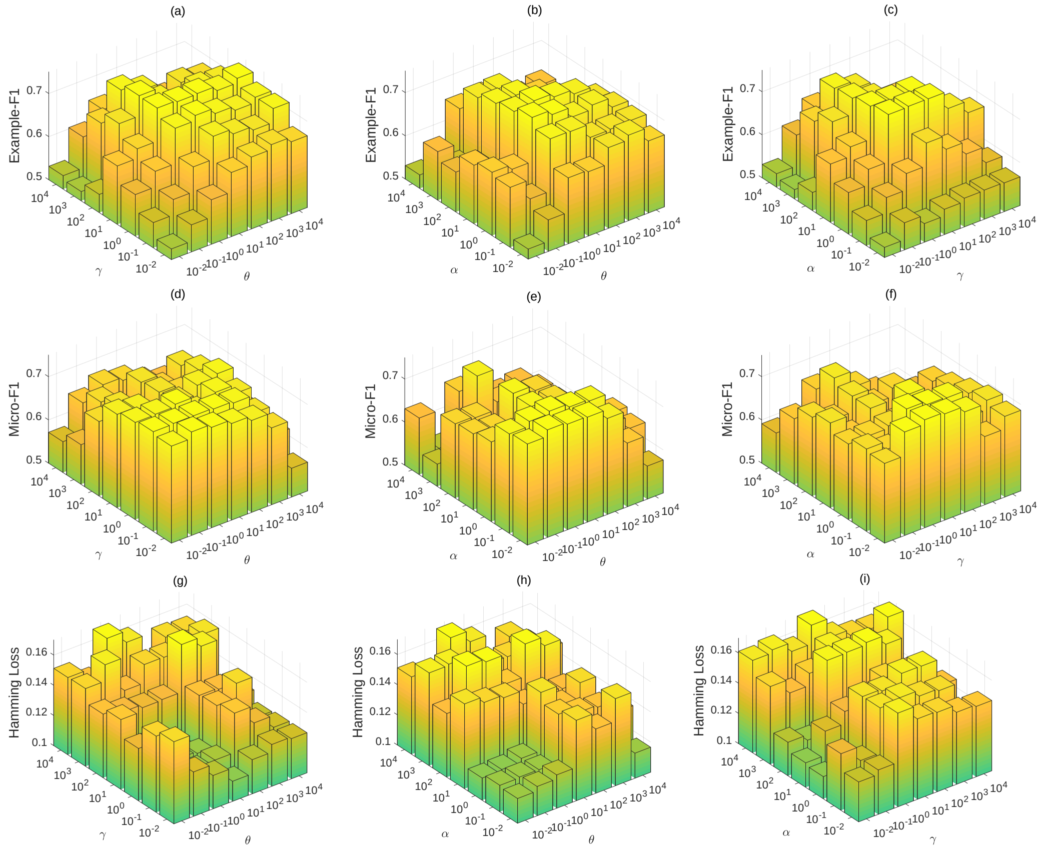

Figure 6. The performance of our proposed method on the Scene dataset through pair-wise comparison for parameters α, γ and θ. (**a**) Example-F1 for (γ, θ) pair, (**b**) Example-F1 for (α, θ) pair, (**c**) Example-F1 for (α, γ) pair, (**d**) Micro-F1 for (γ, θ) pair, (**e**) Micro-F1 for (α, θ) pair, (**f**) Micro-F1 for (α, γ) pair, (**g**) Hamming Los for (γ, θ) pair, (**h**) Hamming Los for (α, θ) pair, and (**i**) Hamming Los for (α, γ) pair.

From Equations (38)–(40), the KKT conditions can be obtained: $E - X + VU^T = 0$, $Z - U0$, $H - S = 0$. This can be used to determine the following convergence conditions: $||E - X + VU^T||_\infty < \varepsilon$, $||Z - U||_\infty < \varepsilon$, $||H - S||_\infty < \varepsilon$. As such, the convergence curves of Figure 7 were drawn on the Scene and Yeast datasets. The observation results in Figure 7 indicate that our method converges quickly within finite iterations.

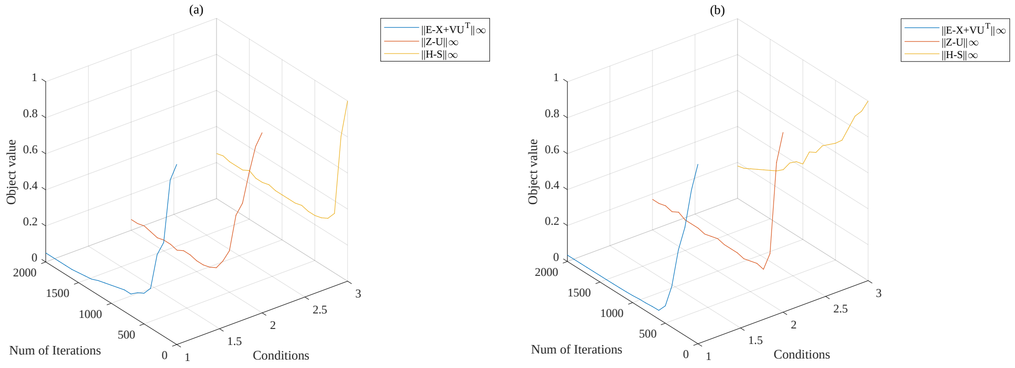

Figure 7. The evolution of the objective values of our proposed method on the Scene (**a**) and Yeast (**b**) datasets.

5. Conclusions

This paper presents a novel approach for ranking support vector machines, which is based on manifold regularization matrix factorization. The proposed method combines feature selection with multi-label learning, creating a unified framework. Within this framework, matrix factorization is introduced to learn low-rank robust subspaces in the input space, thereby enhancing the robustness of data representation in the presence of high levels of noise. Furthermore, by preserving the manifold relationship among low-rank samples, the framework incorporates manifold structure regularization to further improve the robustness of the low-rank representation. Subsequently, using this robust low-rank representation, we extract low-order features that are more resilient and employ them to construct superior classifiers. In other words, these methods enable our classification model to perform better in the presence of original noise. To validate the feasibility of the proposed algorithm, we conducted extensive experiments on real-world multi-label datasets. We compared our algorithm against five commonly used multilabel classification algorithms using multiple evaluation metrics. The experimental results demonstrate that the proposed method outperforms state-of-the-art methods across all evaluation metrics. In the future, we plan to introduce deep learning methods to further enhance multi-label classification models.

To further develop our approach, firstly, different models may be appropriate for different fields. Regardless of the data type, the goal of feature extraction is to find the most representative and informative features in order to improve the performance of machine learning models. The dimensions, processing techniques, and final feature composition that need to be considered vary for different types of data. Therefore, the adaptation of different models on different datasets is a question worthy of further research. Secondly, when dealing with the increasingly large and complex data, the generalization ability and scalability of our model may need to be improved. Chen et al. [32] introduced graph convolutional networks to handle multi label image recognition tasks, utilizing the relationships between labels to improve classification performance. Durand et al. [33] proposed a deep convolutional network to handle this incomplete multi label learning problem. These methods are very inspiring for us, and in the future, we can also introduce deep learning methods to further improve multi label classification models.

Author Contributions: Conceptualization, H.S., Y.Z. and X.S.; Methodology, H.S., Y.Z. and X.S.; Software, Y.Z., E.Q. and Q.Z.; Supervision, H.S. and X.S.; Writing—original draft, H.S., Y.Z., E.Q., Q.Z. and X.S.; Writing—review and editing, H.S., Y.Z. and X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under grant 62172193 and the Yangzhou Industrial Foresight and Common Key Technology Projects under grant SCY2023000087.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://mulan.sourceforge.net/datasets-mlc.html (accessed on 8 July 2023).

Acknowledgments: The authors would like to thank anonymous Reviewers for their valuable reviews.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Schapire, R.E.; Singer, Y. BoosTexter: A boosting-based system for text categorization. *Mach. Learn* **2000**, *39*, 135–168. [CrossRef]
2. Ueda, N.; Saito, K. Parametric mixture models for multi-labeled text. In Proceedings of the NIPS'02: Proceedings of International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 9–14 December 2002; MIT Press: Cambridge, MA, USA, 2002; pp. 721–728.
3. Gao, S.; Wu, W.; Lee, C.H.; Chua, T.S. A MFoM learning approach to robust multiclass multi-label text categorization. In Proceedings of the ICML, Banff, AB, Canada, 4–8 July 2004; pp. 1–8. [CrossRef]

4. Boutell, M.R.; Luo, J.B.; Shen, X.P.; Brown, C.M. Learning multi-label scene classification. *Pattern Recognit.* **2004**, *37*, 1757–1771. [CrossRef]
5. Qi, G.J.; Hua, X.S.; Yong, R.; Tang, J.H.; Tao, M.; Zhang, H.J. Correlative multi-label video annotation. In Proceedings of the ACM International Conference on Multimedia, Augsburg, Germany, 25–29 September 2007; pp. 17–26. [CrossRef]
6. Trohidis, K.; Tsoumakas, G.; Kalliris, G. Multi-label classification of music by emotion. *EURASIP J. Audio Speech Music. Process.* **2011**, *4*, 1–9. [CrossRef]
7. Liu, Y.; Wen, K.; Gao, Q.; Gao, X.; Nie, F. SVM based multi-label learning with missing labels for image annotation. *Pattern Recognit.* **2018**, *78*, 307–317. [CrossRef]
8. Li, J.; Zhang, C.; Zhou, J.T.; Fu, H.; Xia, S.; Hu, Q. Deep-LIFT: Deep label-specific feature learning for image annotation. *IEEE Trans. Cybern.* **2021**, *52*, 7732–7741. [CrossRef]
9. Xing, Y.; Yu, G.; Domeniconi, C. Multi-Label Co-Training. In Proceedings of Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 2882–2888. Available online: https://dl.acm.org/doi/abs/10.5555/3304889.3305061 (accessed on 8 July 2023).
10. Zhang, M.L.; Li, Y.K.; Yang, H.; Liu, X.Y. Towards class-imbalance aware multi-label learning. *IEEE Trans. Cybern.* **2020**, *52*, 4459–4471. [CrossRef]
11. Cortes, C.; Mohri, M. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems: Proceedings of the 2004 Conference*; MIT Press: Cambridge, MA, USA, 2004; pp. 313–320.
12. Wu, X.Z.; Zhou, Z.H. A unified view of multi-label performance measures. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3780–3788.
13. Wu, B.; Lyu, S.; Hu, B.; Ji, Q. Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognit.* **2015**, *48*, 2279–2289. [CrossRef]
14. Cevikalp, H.; Benligiray, B.; Gerek, O.N. Semi-supervised robust deep neural networks for multi-label image classification. *Pattern Recognit.* **2020**, *100*, 107164. [CrossRef]
15. Kuang, D.; Ding, C.; Park, H. Symmetric nonnegative matrix factorization for graph clustering. In Proceedings of the SIAM International Conference on Data Mining, Anaheim, CA, USA, 26–28 April 2012; pp. 106–117. [CrossRef]
16. Izenman, A. Introduction to manifold learning. *Comput. Stat.* **2012**, *4*, 439–446. [CrossRef]
17. Fuernkranz, J.; Huellermeier, E.; Mencia, E.L.; Brinker, K. Multilabel classification via calibrated label ranking. *Mach. Learn.* **2008**, *73*, 133–153. [CrossRef]
18. Nasrabadi, N.M.; King, R.A. Image coding using vector quantization: A review. *IEEE Trans. Commun.* **1988**, *36*, 957–971. [CrossRef]
19. Elisseeff, A.; Weston, J. A kernel method for multi-labelled classification. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2001; pp. 681–687.
20. Zhang, M.L.; Zhou, Z.H. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.* **2007**, *40*, 2038–2048. [CrossRef]
21. Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. Classifier chains for multi-label classification. *Mach. Learn.* **2011**, *85*, 333–359. [CrossRef]
22. Zhang, M.L.; Zhou, Z.H. Multi-label neural networks with applications to functional genomics and text categorization. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 1338–1351. [CrossRef]
23. Jiang, A.; Wang, C.; Zhu, Y. Calibrated Rank-SVM for multi-label image categorization. In Proceedings of the IEEE International Joint Conference on Neural Networks, Hong Kong, China, 1–8 June 2008; pp. 1450–1455. [CrossRef]
24. Xu, J. An efficient multi-label support vector machine with a zero label. *Expert Syst. Appl.* **2012**, *39*, 4796–4804. [CrossRef]
25. Yang, J.; Kai, Y.; Gong, Y.; Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1794–1801. [CrossRef]
26. Wright, J.; Yang, A.Y.; Ganesh, A.; Sastry, S.S.; Ma, Y. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 210–227. [CrossRef] [PubMed]
27. Ouyang, T.H. Feature learning for stacked elm via low-rank matrix factorization. *Neurocomputing* **2021**, *448*, 82–93. [CrossRef]
28. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2006**, *68*, 49–67. [CrossRef]
29. Wu, G.Q.; Tian, Y.J.; Liu, D.L. Cost-sensitive multi-label learning with positive and negative label pairwise correlations. *Neural Netw.* **2018**, *108*, 411–423. [CrossRef]
30. Wu, G.Q.; Li, C.X.; Xu, K.; Zhu, J. Rethinking and Reweighting the Univariate Losses for Multi-Label Ranking: Consistency and Generalization. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; pp. 14332–14344. Available online: https://proceedings.neurips.cc/paper/2021/hash/781397bc0630d47ab531ea850bddcf63-Abstract.html (accessed on 8 July 2023).
31. Wu, G.Q.; Zheng, R.B.; Tian, Y.J.; Liu, D.L. Joint ranking SVM and binary relevance with robust low-rank learning for multi-label classification. *Neural Netw.* **2020**, *122*, 24–39. [CrossRef]

32. Chen, Z.M.; Wei, X.S.; Wang, P.; Guo, Y.W. Multi-Label Image Recognition with Graph Convolutional Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5172–5181. [CrossRef]
33. Durand, T.; Mehrasa, N.; Mori, G. Learning a Deep ConvNet for Multi-label Classification with Partial Labels. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 647–657. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.