A novel multi-label learning approach for missing label and noisy label

Kai Liu[1], Anhui Tan[1,1,], Shen-ming Gu[1]

[a]School of Information Engineering, Zhejiang Ocean University, Zhoushan, Zhejiang 316022, PR China
[b]School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, Shanxi, PR China

Abstract

Weakly-supervised multi-label learning has brought about the widespread attention recently. Most of the existing methods solve this problem by assuming that the label information is ideal. However, in some practical applications, partial labels of each instance can only be obtained. In addition the labels obtained may be noisy, which will reduce the robustness of the learning model. In this paper, we propose a novel approach named a novel multi-label learning approach for missing label and noisy label(MLNL) to overcome the above problems. On the one hand, we introduce a label correlation matrix to explicitly utilize the global structure information from multi-label data. On the other hand, we propose a new label manifold regularizer to capture noisy labels to further improve the performance of the method. Finally, the performance of the method is further improved by considering the positive and negative information between the feature similarity and label similarity. The extensive experimental results show that our algorithm MLNL has comparable performance compared with some state-of-the-art methods when tested on bennhmark multi-label date sets.

Keywords: Multi-label; noisy label; feature collaboration; label collaboration; optimization objective

1. Introduction

With the rapid development of Internet technology, the traditional single label learning gradually shows its limitations. In traditional single-label learning, each instance is associated with a single label. However, in the real application scenario, most of the single instance is matched with multiple labels. Therefore, multi-label learning has aroused the interest of many researchers. At the same time, due to of the influence of technology and methods, the extracted database inevitably has default values and noise values. Aiming at such incomplete and inaccurate data, the traditional supervised multi-label learning model lacks the ability of accurate learning. Furthermore, some scholars put forward a weakly supervised multi-label learning model[Cabral R , De l T F , Costeira J P , et al. Matrix Completion for Weakly-Supervised Multi-Label Image Classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1):121-35.][Sun L , Ye P , Lyu G , et al. Weakly-Supervised Multi-Label Learning with Noisy Features and Incomplete Labels[J]. Neurocomputing, 2020, 413.][Zhou Z H . A Brief Introduction to Weakly Supervised Learning[J]. National Science Review, 2018(1):1], which provides an effective method to solve such problem with missing and noisy labels. So far, it has been applied in many fields, such as image classification, activity recognition and text classification[Cabral R , Torre F , Costeira J P , et al. Matrix Completion for Weakly-Supervised Multi-Label Image Classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.][Adeli-Mosabbeb E , Cabral R , Torre F , et al. Multi-label Discriminative Weakly-Supervised Human Activity Recognition and Localization[J]. 2014.][Mercan C , Aksoy S , Mercan E , et al. Multi-Instance Multi-Label Learning for Multi-Class Classification of Whole Slide Breast Histopathology Images[J]. IEEE Trans Med Imaging, 2017, PP(99):1-1.].

[*] Corresponding author.

In multi-label learning, the traditional multi label learning algorithm is designed on the premise that the label matrix of training data is complete and correct, such as [Nie F , Huang H , Xiao C , et al. Efficient and Robust Feature Selection via Joint ?2, 1-Norms Minimization[C] International Conference on Neural Information Processing Systems. Curran Associates Inc. 2010.] and MDFS[Zhang J , Luo Z , Li C , et al. Manifold regularized discriminative feature selection for multi-label learning[J]. Pattern Recognition, 2019, 95:136-150.]. However, humans may neglect some labels they don't know or aren't interested in, which will lead to missing labels in the training set. In order to overcome this problem, scholars try to recover missing labels by label correlation. For example, in the basic of label correlated, Some people assume that label correlation matrix and feature-lable mapping matrix put into a unified linear dependent structure and propose low rank hypothesis and sparse hypothesis[Zhu Y , Kwok J T , Zhou Z H . Multi-Label Learning with Global and Local Label Correlation[J]. IEEE Transactions on Knowledge Data Engineering, 2017:1-1][Hao-Chen Dong, Yu-Feng Li, Zhi-Hua Zhou:Learning From Semi-Supervised Weak-Label Data. AAAI 2018: 2926-2933]. A common approach to use the low rank hypothesis is through the nuclear-norm regularization term. However, the above methods are based on the assumption that the obtained label information is ideal, and the problem of noise labels in the label information is not taken into account. Obviously, the presence of noise labels will participate in the multi-label training and seriously affect the final multi-label learning model. for noise label information, the existing strategies can be roughly divided into two types: some people adopt a unified learning structure, estimate the confidence of each label from multi-label learning, and then incorporate the estimated confidence matrix into the optimization process induced by the model[L. Sun, S. Feng, T. Wang, C. Lang, Y. Jin, Partial multi-label learning via low-rank and sparse decomposition, in: AAAI Conference on Artificial Intelligence, 2019, pp. 5016?5023]; Another approach is to divide the training process into two steps: first, select labels with high confidence from the label set, and then use multi-label learning structure training from these labels with high confidence[J. Fang, M. Zhang, Partial multi-label learning via credible label elicitation, in: AAAI Conference on Artificial Intelligence, 2019, pp. 3518-3525.].Some scholars also proposed a multi-label learning algorithm aiming at incomplete data and noise feature information at the same time. such as, Sun[Sun L , Lyu G , Feng S , et al. Beyond missing: weakly-supervised multi-label learning with incomplete and noisy labels[J]. Applied Intelligence, 2021, 51(3):1-13] used low rank hypothesis and sparse constraints to sort sensitive labels by enrich the missing labels and remove the noisy labels simultaneously. sun[Sun L , Ye P , Lyu G , et al. Weakly-Supervised Multi-Label Learning with Noisy Features and Incomplete Labels[J]. Neurocomputing, 2020, 413]and Gengyu[Gl A , Sf A , Yl A . Noisy label tolerance: A new perspective of Partial Multi-Label Learning[J]. Information Sciences, 2021, 543:454-466] capture the desired feature information by decomposing the observed feature matrix into an ideal feature matrix and an outlier matrix. nevertheless, they didn't consider the possibility of noise labels when labels are missing, and paid attention to labels with high feature similarity and high label similarity when building models, but ignored the negative information between features and labels, that is, the situation of high (low) feature similarity and low (high) label similarity. To alleviate this problem, this paper proposes a new multi-label learning method based on missing and noisy labels, called"A novel multi-label learning approach for missing label and noisy label"(MLNL), which simultaneously recovers missing labels, excludes noisy labels, trains the linear model, expores and utilizes label correlation. Firstly, MLNL learns a linear predictor about the ground-truth label matrix, and limits the low rank of the ground-truth label matrix and the sparsity of the noise label in the prediction. At the same time, the label correlation matrix is introduced to explain the correlation of multi-label data. Then, using the feature similarity and label similarity of instances, the inner product similarity of potential label vector is calculated, so that the inner product similarity of potential label vector and the feature of samples are consistent with the semantic similarity of labels. Finally, the predictor and the above constraints are integrated into a unified linear model, and this paper develops an optimization procedure to solve the model. The motivation of this article is summarized as folllows.

We propose a new multi label learning method to deal with missing labels and noisy labels simultaneously.

We introduce MLNL to induce a credible multi label classifier by using the label and feature information of multi-label data. The existing multi-label learning methods utilize label and feature information separately, or ignore the utilization of negative information. Two candidate labels with high (low) feature similarity but low (high) semantic similarity are selected.

MLMN unifies the predictor training and potential label matrix exploration of multi label data in a unified linear model, and introduces an alternative optimization program to optimize the predictor and potential label matrix in a mutually beneficial way.

The rest of this article is organized as follows. Firstly, we briefly discuss the related work of multi-label learning with incomplete label information and noise label information.Then, we give the technical details of the proposed algorithm and describe the corresponding optimization process. Then the comparative experiment is carried out and the further experimental analysis is carried out. Finally, we summarize the whole paper.

2. Related work

2.1. Multi-label learning with missing label

It is not appropriate to assume that all labels are observed. Thus many scholars have proposed relevant algorithms for missing labels, such as Wu[Wu B , Lyu S , Ghanem B . ML-MG: Multi-label Learning with Missing Labels Using a Mixed Graph[C] IEEE International Conference on Computer Vision (ICCV). IEEE, 2015.]propose a unified model of label dependencies by constructing a mixed graph, which jointly incorporates instance-level similarity and class co-occurrence as undirected edges as well as semantic label hierarchy as directed edges . Ma[Ma Z , Chen S . Expand globally, shrink locally: Discriminant multi-label learning with missing labels[J]. Pattern Recognition, 2021, 111:107675]impose the low-rank structures on all the predictions of instances from the same labels, and a maximally separated structure on the predictions of instances from different labels. Huang[J Huang, Qin F , Zheng X , et al. Improving Multi-Label Classification with Missing Labels by Learning Label-Specific Features[J]. Information Sciences, 2019]propose a new supplementary label matrix is augmented from the in complete label matrix by learning high-order label correlations.

In order to learn accurately from multi-label data with default values and noise labels, it is necessary to solve the problem of how to deal with incomplete label and noise information.On the one hand, for the problem of incomplete labels, the existing algorithms are mainly divided into four strategies: first, the correct labels are simply regarded as negative labels, such as[Bucak S, Jin R, Jain A Multi-label learning with incomplete class assignments. In: IEEE Conference on computer vision and pattern recognition, 2011, pp. 2801-2808];Secondly, missing tags are regarded as potential variables and embedded in probability models, such as [Zhang J , Li S , Jiang M , et al. Learning From Weakly Labeled Data Based on Manifold Regularized Sparse Model[J]. IEEE Transactions on Cybernetics, 2020, PP(99)] and Bayesian networks[Kapoor A , Jain P , Viswanathan R . Multilabel Classification using Bayesian Compressed Sensing.[J]. Advances in Neural Information Processing Systems, 2012][Sucar L E , Bielza C , Morales E F , et al. Multi-label classification with Bayesian network-based chain classifiers[J]. Pattern Recognition Letters, 2014, 41(MAY 1):14-22].Thirdly, the missing label is regarded as an independent state, and all the labels are divided into three categories: positive label, negative label and missing label[Wu B, Jia F, Liu W, Ghanem B, Lyu S (2018) Multi-label learning with missing labels using mixed dependency graphs. Int J Comput Vis 126(8):875?896].Fourthly, based on the method of matrix incompletion, the information matrix is used to restore the missing items of label matrix[Chang X , Tao D , Chao X . Robust Extreme Multi-label Learning[C] Acm Sigkdd International Conference on Knowledge Discovery Data Mining. ACM, 2016][Zhang J , Li S , Jiang M , et al. Learning From Weakly Labeled Data Based on Manifold Regularized Sparse Model[J]. IEEE Transactions on Cybernetics, 2020, PP(99)].

2.2. Multi-label learning with noisy label

Multi label learning with noise label (also known as partial multi-label learning (PML) is a new learning framework, which deals with the scenario where multiple candidate labels are assigned to a case with only partial validity.Li[Li Z , Gan Z , Zhang B , et al. Semi-Supervised Noisy Label Learning for Chinese Medical Named Entity Recognition[J]. Data Intelligence, 2021:1-10]propose a semi-supervised hand segmentation framework based on the optimized noisy masks and a small number of labeled data. wang[Wang

Z , Jiang J , Han B , et al. SemiNLL: A Framework of Noisy-Label Learning by Semi-Supervised Learning[J]. 2020.] propose a versatile framework that combines specific sample selection strategies and specific semi-supervised learning models in an end-to-end manner. Liu[Liu T , Tao D . Classification with Noisy Labels by Importance Reweighting[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2016, 38(3):447-461] prove that any surrogate loss function can be used for classification with noisy labels by using importance reweighting, with consistency assurance that the label noise does not ultimately hinder the search for the optimal classifier of the noise-free sample. xie[Xie M K , Huang S J . Partial Multi-Label Learning with Noisy Label Identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, PP(99).]propose the MIPML method that the multi-label classifier and noisy label identifier are learned jointly under the supervision of the observed noise-corrupted label matrix.

3. The proposed model

3.1. Notations

In order to formulize the learning of partial label data, some important use marks in this paper are explained as follows.In multi-label learning, given weakly multi-lable dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$ with n examples. we denote the feature matrix $X = [x_1, x_2, \ldots, x_n] \in R^{d\times n}$, where x_i is d dimension column. Let $Y = [y_1, y_2, \ldots, y_n]^T \in \{0, 1\}^{n\times q}$ be the label matrix, and $y_i (1 \leq i \leq n) \in R^q$ is the ground-truth label vector of x_i. For $y_{ij} \in Y$, $y_{ij} = 1$ represents the jth label is relevent to the ith instance, and $y_i j = 0$ represents the ith instance is nothing to do with the jth label or the jth label is missing. The Frobenius norm of the matrix $A = (a_{ij})_{d\times p}$ is defined as $\|A\|_F = \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{q} a_{ij}^2}$. The nuclear norm of A is defined as the sum of its singular values, which is also equivalent to the $tr\left(\sqrt{A^T A}\right)$. The $l_2, 1$-norm of A is defined as $\|W\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{q} w_{ij}^2}$.

3.2. The basic framework

For gaining such a linear model for Weakly-supervised multi-label learning, we can generalize the learning problem as the following optimization formulation:

$$\min_{W} \Phi(W, b) + \lambda_1 \Re(W) \tag{1}$$

where $\Phi(W, b)$ is a loss function, $\Re(W)$ is the regularizad term to control the complexity of the model. λ_1 is the trading off parameters with non-negative value.

The loss function $\Phi(W, b)$ can be applied in various ways. Here we choose the robust loss function for weakly multi-label learning considering implicity and efficiency, as follows:$\Phi(W, b) = \sum_{i=1}^{q} \sum_{i=1}^{n} \left(x_i w_k + b_k - y_{ij}\right)^2$. Therefore, the loss function can be rewrited as:

$$\Phi(W, b) = \left\|X^T W - 1_n b - Y\right\|_F^2 \tag{2}$$

where $W = \left[w_1, w_2, \ldots, w_q\right] \in R^{d\times q}$ is the coefficient matrix. $b \in R^{1\times n}$ is the bias term, and 1_n is n dimension column with all elements are one.

In consideration that making W sparse, we implement $l_2, 1$-norm regularization into our linear model

$$\Re(W) = \|W\|_{2,1} \tag{3}$$

Thus,the w_{ij} reflects that the ith feature associated with the jth label. When the value of w_{ij} is large, the ith feature has strongly discriminability to the jth label, On the contrary, even if it tends to zero, which means they have nothing to do with each other.

In our problem, we assume that there is a small amount of noise information in the label matrix. In order to make full use of the ideal information of the label, we decompose the label matrix into two parts to remove the influence of noise information, i.e. $Y = Q + N$, where $Q \in \{0, 1\}^{n\times q}$ is the ideal label matrix and $N \in \{0, 1\}^{n\times q}$ is the noise matrix with abnormal label values. Considering the label correlation, we limit the ideal label matrix Q to low rank. Because the rank function is difficult to solve, the kernel norm is used to replace the rank function. In addition, we assume that the outliers are sparse relative to the observed label matrix Y, so We utilize l_1 norm to restrict noise matrix N to be a sparse matrix. Therefore, the proposed model can be solved by the following objective functions:

$$\begin{array}{l} \min_{W,Q,N} \left\| X^T W - 1_n b - P \right\|_F^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \|Q\|_* + \lambda_3 \|N\|_1 \\ \text{s.t. } Y = Q + N \end{array} \tag{4}$$

where λ_1, λ_2, λ_3 are tradeoff parameters to balance the model.

As noted above, despite of the labels in the ideal label matrix Q are correct, there still some label are missing in Q. Ling et al [?] proved that the multi label output space can be mathematically decomposed into the product of two low dimensional matrices, which encourages most of the structures in the original low dimensional output label matrix to be interpreted and restored. Therefore, we adopt a linear self recovery model $P = CQ$ to reconstruct Q. Here, $C \in R^{n \times n}$ is label coefficient matrix, and the element C_{ij} represents the similarity between the $i - th$ label and the $j - th$ label. A meaningful matrix self recovery model should automatically complete the unknown labels in the ideal label matrix Q without affecting the majority of observed labels. Finally, we propose the following objective function to learn the self recovery model:

$$\begin{array}{l} \min_P \lambda_4 \|P - CQ\|_F^2 + \lambda_5 \|C\|_* \\ \text{s.t. } \sum_{j=1}^n C_{ij} = 1, C \geq 0, P \geq 0 \end{array} \tag{5}$$

where λ_4, λ_5 are tradeoff parameters to balance the model. Here we assume the label coefficient matrix C be low-rank which exploit the latent label correlations among different labels. the first constraint restricts the value to be in the range of $[0, 1]$, and the sum of them equal to 1.

Most of the existing partial label learning (PML) problems are based on similar (different) samples with the same (different) label assignment in feature space. They use the similarity of feature space for Manifold Regularization label learning. Some scholars think that the label space is low rank based on the knowledge of latent label space, and the related lables of a sample are hidden in the latent lable space. Although these works use feature information to label training samples to some extent, they still consider high feature correlation and potential tag correlation. In the actual partial label learning (PML) problem, if the potential real labels of two samples do not overlap, then the two samples should also have no duplicate real labels. Therefore, the correlation between the two samples is not significant. In a word, the existing methods are not accurate enough for the samples with high feature correlation but low potential lable correlation. To solve this problem, we formulate the following formula:

$$\lambda_6 \sum_{i \neq j}^n \left(S_{ij} C_{ij} - p_{i,} p_{j_r} \right)^2 \tag{6}$$

where λ_6 are tradeoff parameters to balance the model. $S \in R^{n \times n}$ is the sample similarity matrix about the feature space, and each element S_{ij} represents the sample similarity between sample x_i and sample x_j. The calculation formula is as follows:

$$S_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2} \right) & \text{if } x_i \in \mathcal{N}_k\left(x_j\right) \text{ or } x_j \in \mathcal{N}_k\left(x_i\right) \\ & 0 \text{ otherwise} \end{cases} \tag{7}$$

where σ is a parameter. $\mathcal{N}_k(x_i)$ denotes the set of k largest nearest neighbors of x_i in the feature space, and Euclidean distance is used to find k nearest neighbors.

From the above we can find that: If two samples have larger S_{ij} and C_{ij} at the same time, p_i and p_j should be close to each other; If the value of S_{ij} is large but the value of C_{ij} is small or or vice versa, then p_i and p_j may still have some repetitions; If two samples have small S_{ij} and C_{ij} at the same time,p_i and p_j should have little or no coincidence.

By combining problem (4),(5),(6), and equation (8), the final optimization problem can be further rewritten as follows:

$$\begin{array}{l}\min_{W,Q,N,P,C,d} \left\|X^T W - 1_n b - P\right\|_F^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \|Q\|_* + \lambda_3 \|N\|_1 + \lambda_4 \|P - CQ\|_F^2 \\ +\lambda_5 \|C\|_* + \lambda_6 \sum_{i \neq j}^n \left(S_{ij} C_{ij} - p_{i,p} p_{j,}\right)^2 \\ \text{s.t. } Y = Q + N, \sum^n C_{ij} = 1, C \geq 0, P \geq 0\end{array} \tag{8}$$

where $\lambda_l (l = 1, 2, 3, 4, 5, 6)$ are tradeoff parameters to balance the model.

4. Optimization and solution

For the convenience of calculation, the original objective function is optimized as follows:

$$\begin{array}{l}\min_{W,Q,N,P,C,d} \left\|X^T W - 1_n b - P\right\|_F^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \|Q\|_* + \lambda_3 \|N\|_1 + \lambda_4 \|P - CQ\|_F^2 \\ +\lambda_5 \|C\|_* + \lambda_6 \left\|H \odot \left(S \odot C - PP^T\right)\right\|_F^2 \\ \text{s.t. } Y = Q + N, C1_n = 1_n, C \geq 0, P \geq 0\end{array} \tag{9}$$

where $H \in R^{n \times n}$, if $i = j$, then $H_{ij} = 0$, otherwise $H_{ij} = 1$. $\odot$ is the Hadamard product. 1_n is the n-dimensional column vector whose elements are all 1.

By using Lagrange multiplier method[Shimizu A . Lagrange-Multiplier Method[J]. Journal of the Japan Society of Mechanical Engineers, 2009, 112(5):987-992], the above formula can be rewritten into the form of augmented Lagrange function as follows:

$$\begin{array}{l}\min_{W,Q,N,P,C,d} \left\|X^T W - 1_n b - P\right\|_F^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \|Q\|_* + \lambda_3 \|N\|_1 + \lambda_4 \|P - CQ\|_F^2 \\ +\lambda_5 \|C\|_* + \lambda_6 \left\|H \odot \left(S \odot C - PP^T\right)\right\|_F^2 \\ + < Z_1, Y - Q - N > + \frac{\mu_1}{2} \|Y - Q - N\|_F^2 \\ + < Z_2, C|_n - 1_n > + \frac{\mu_2}{2} \|C1_n - 1_n\|_F^2\end{array} \tag{10}$$

As the model is linear, the above formula can be further rewritten by LADMAP method[Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in: Advances in Neural Information Processing Systems, 2011, pp. 612-620] as follows:

$$\begin{array}{l}\min_{W,Q,N,P,C,d} \left\|X^T W - 1_n b - P\right\|_F^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \|Q\|_* + \lambda_3 \|N\|_1 + \lambda_4 \|P - CQ\|_F^2 \\ +\lambda_5 \|C\|_* + \lambda_6 \left\|H \odot \left(S \odot C - PP^T\right)\right\|_F^2 \\ + \frac{\mu_1}{2} \left\|Y - Q - N + \frac{Z_1}{\mu_1}\right\|_F^2 + \frac{\mu_2}{2} \left\|C1_n - 1_n + \frac{Z_2}{\mu_2}\right\|_F^2\end{array} \tag{11}$$

where $Z_1 \in R^{n \times q}$, $Z_2 \in R^n$ represent the Lagrange multiplier matrix, and μ_1, μ_2 is the penalty parameter.

4.1. Update b

First, fixd W, P, C, Q, N and updated d, d can be obtained by solving the following subproblems:

$$\min_b \left\|X^T W - 1_n b - P\right\|_F^2 \tag{12}$$

and we can get the closed form solution by setting its derivative to zero:

$$b = \frac{1}{n} \left(1_n^T X^T W - 1_n^T P\right) \tag{13}$$

4.2. Update W

Similarly, fixd d, P, C, Q, N and updated W, W can be obtained by solving the following subproblems:

$$\min_W \left\|X^T W - 1_n b - P\right\|_F^2 + \lambda_1 \|W\|_{2,1} \tag{14}$$

Then the gradient of W is

$$\nabla W = 2X\left(X^T W - 1_n b - P\right) + \lambda_1 \Sigma W \tag{15}$$

where Σ is defined as

$$\Sigma = \begin{pmatrix} \frac{1}{\|w_1\|_2} & & & \\ & \frac{1}{\|w_2\|_2} & & \\ & & \cdots & \\ & & & \frac{1}{\|w_d\|_2} \end{pmatrix}$$

Let the equation(16) be equal to zero,

$$W = \left(2XX^T + \lambda_1 \Sigma\right)^{-1} 2X\left(1_n b + P\right) \tag{16}$$

4.3. Update P

Similarly, fixd d, W, C, Q, N and updated P, P can be obtained by solving the following subproblems:

$$\min_P \left\|X^T W - 1_n b - P\right\|_F^2 + \lambda_6 \left\|H \odot \left(S \odot C - PP^T\right)\right\|_F^2 + \lambda_4 \|P - CQ|_F^2 \tag{17}$$

Then the gradient of P is

$$\begin{aligned} \nabla P = &\left(P - X^T W + 1_n b\right) + \lambda_6 H \odot \left((S \odot C) - PP^T\right) P \\ &+ \lambda_6 H^T \odot \left(\left(S^T \odot C^T\right) - PP^T\right) P + \lambda_4 (P - CQ) \end{aligned} \tag{18}$$

We can use the KKT condition for the nonnegativity of P as:

$$\begin{aligned} &(((P - X^T W + 1_n b) + \lambda_6 (H \odot ((S \odot C) - PP^T)P \\ &+ \lambda_6 (H^T \odot (S^T \odot C^T) - PP^T)P + \lambda_4 (P - CQ))_{ij} P_{ij} = 0 \end{aligned} \tag{19}$$

let $(X^T W = (X^T W^+ - X^T W^-)$, then the above formula can be rewritten as

$$\begin{aligned} &((P - X^T W^+ - X^T W^- + 1_n b) + \lambda_6 (H \odot ((S \odot C) - PP^T)P \\ &+ \lambda_6 (H^T \odot (S^T \odot C^T) - PP^T)P + \lambda_4 (P - CQ))_{ij} P_{ij} = 0 \end{aligned} \tag{20}$$

Furthermore, P can be updated by the following formula:

$$P_{ij} = P_{ij} \sqrt{\frac{\left(X^T W^+ + \lambda_7 \left(H \odot PP^T\right) P + \lambda_7 \left(H^T \odot PP^T\right) P + \lambda_4 CQ\right)_{ij}}{\left(P + X^T W^- + 1_n b + \lambda_7 H \odot S \odot C + + \lambda_7 H^T \odot S^T \odot C^T + \lambda_4 P\right)_{ij}}} \tag{21}$$

4.4. Update C

Similarly, fixd d, W, P, Q, N and updated C, C can be obtained by solving the following subproblems:

$$\min_C \lambda_6 \left\|H \odot \left(S \odot C - PP^T\right)\right\|_F^2 + \lambda_4 \|P - CQ|_F^2 + \frac{\mu_2}{2} \left\|C_n - 1_n + \frac{z_2}{\mu_2}\right\|_F^2 + \lambda_5 \|C\|_* \tag{22}$$

Derivation of the first term and the second term with respect to C:

$$\nabla C = 2\lambda_6 H \odot S^T \odot \left(S \odot C - PP^T\right) + 2\lambda_4 (CQ - P) Q^T + \mu_2 \left(C 1_n - 1_n\right) 1_n^T \tag{23}$$

$$\begin{aligned}
&\|\nabla C_1 - \nabla C_2\|_F^2 \\
&= \left\|2\lambda_6 H \odot S^T \odot S \odot \Delta C + 2\lambda_4 \Delta C Q Q^T + \mu_2 \Delta C 1_n 1_n^T\right\|_F^2 \\
&\leq 3\left\|2\lambda_6 H \odot S^T \odot S \odot \Delta C\right\|_F^2 + 3\left\|2\lambda_4 \Delta C Q Q^T\right\|_F^2 + \left\|\mu_2 \Delta C 1_n 1_n^T\right\|_F^2 \\
&\leq 12\lambda_6^2 \max\left\{s_{ij}^2\right\}\|\Delta C\|_F^2 + 12\lambda_4^2\left\|Q Q^T\right\|_2^2\|\Delta C\|_F^2 + 3\mu_2^2\left\|1_n 1_n^T\right\|_2^2\|\Delta C\|_F^2 \\
&= \left(12\lambda_6^2 \max\left\{s_{ij}^2\right\} + 12\lambda_4^2\left\|Q Q^T\right\|_2^2 + 3n\mu_2^2\right)\|\Delta C\|_F^2
\end{aligned} \tag{24}$$

So Lipschitz constant can be obtained by the following formula:

$$L_C = 12\lambda_7^2 \max\left\{s_{ij}^2\right\} + 12\lambda_4^2\left\|Q Q^T\right\|_2^2 + n\mu_2^2 \tag{25}$$

In the accelerated proximal gradient method[Ji S, Ye J An accelerated gradient method for trace norm minimization. In: International conference on machine learning, 2009, pp 457-464], the accelerated proximal gradient iterates as follows:

$$C^k = C^k + \frac{\alpha_{k-1} - 1}{\alpha_k}(C_k - C_{k-1}) \tag{26}$$

$$C^t = \mathcal{D}_{\frac{\lambda_5}{L_C}}(C^k - \frac{1}{L_C}\nabla C) \tag{27}$$

where $\mathcal{D}$ is a singular value soft threshold operator[Donoho D , Gavish M . Minimax risk of matrix denoising by singular value thresholding[J]. Annals of Statistics, 2014, 42(6):2413-2440].

4.5. Update Q

Similarly, fixd d, W, P, C, N and updated Q, Q can be obtained by solving the following subproblems:

$$\min_Q \lambda_4 \left\|P - CQ\right|_F^2 + \frac{\mu_1}{2}\left\| Y - Q - N + \frac{Z_1}{\mu_1} \right\|_F^2 + \lambda_2\| Q\Big\|_* \tag{28}$$

Derivation of the first term and the second term with respect to Q:

$$F_Q = \lambda_4 C^T(P - CQ) - \mu_1(Q + N - Y) - Z_1 \tag{29}$$

Then,

$$Q^{k+1} = \mathcal{D}_{\frac{\lambda_2}{\beta}}\left(Z^k - F_{z^k}/\alpha\right) \tag{30}$$

where $\mathcal{D}$ is a singular value soft threshold operator, $\beta = \frac{(2\lambda_4+\mu_1)\tau_Z}{2}$, $\tau_Z > \rho\left(CC^T\right)$, and $\rho\left(CC^T\right)$ is the spectral radius of CC^T.

4.6. Update N

Similarly, fixd d, W, P, C, Q and updated N, N can be obtained by solving the following subproblems:

$$\min_N \lambda_3\|N\|_1 + \frac{\mu_1}{2}\left\|Y - Q - N + \frac{Z_1}{\mu_1}\right\|_F^2 \tag{31}$$

The above formula can be used to obtain the closed solution by the soft threshold operator. Therefore, N can be updated by:

$$N^{k+1} = \mathcal{S}_{\frac{\lambda_3}{\mu_1}}\left[Y - Q^k + \frac{Z_1^k}{\mu_1^k}\right] \tag{32}$$

For any $x \in R$ and $\varepsilon > 0$, the soft threshold operator[Pal P , Vaidyanathan P P . Soft-thresholding for spectrum sensing with coprime samplers[J]. IEEE, 2014] is defined as follows:

$$\mathcal{S}_{\varepsilon}[x] = \begin{cases} x - \varepsilon & x > \varepsilon \\ x + \varepsilon & x < -\varepsilon \\ 0 & else \end{cases} \tag{33}$$

Finally, we update the Lagrange multiplier matrix Z_1, Z_2 and penalty parameter μ_1,μ_2 by the following formula:

$$\begin{aligned} Z_1^{k+1} &= Z_1^k + \mu_1^{k+1}(Y - Q - N) \\ Z_2^{k+1} &= Z_2^k + \mu_2^{k+1}(C1_n - 1_n) \\ \mu_1^{k+1} &= min(\mu_{max}, \rho\mu_1^{k+1}) \\ \mu_2^{k+1} &= min(\mu_{max}, \rho\mu_2^{k+1}) \end{aligned}$$

According to the above analysis, we present a detailed algorithm for A novel multi-label learning approach for missing label and noisy label as shown in Algorithm 1.

Algorithm 1 An algorithm for computing the sample similarity matrix in the label space of a multi-label data set

```
Input:   training instance set $X \in R^{d\times n}$, training label set $Y \in 0,1^{n\times q}$, parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, $\lambda_5$, $\lambda_6$
Output:  The coefficient matrix $W \in R^{d\times q}$
Initialization
Compute the sample similarity matrix $S$ in Eq.(7);$W = 1, P = 1, C = C_0 = 1, Q = 1, N = 0$, $Z_1 = Z_2 = 0$, $\mu_1$, $\mu_2 > 0$
repeat
compute the diagonal matrix $\Sigma$ as:
    $\Sigma = \begin{pmatrix} \frac{1}{\|w_1\|_2} & & & \\ & \frac{1}{\|w_2\|_2} & & \\ & & \cdots & \\ & & & \frac{1}{\|w_d\|_2} \end{pmatrix}$;
compute $L_f$ according to Eq.(26);
Update the optimal $b$ by Eq.(14);
Update the optimal $W$ by Eq.(17);
Update the optimal $P$ by Eq.(22);
    repeat
        Update $Z_2^{k+1}$ by $Z_2^{k+1} = Z_2^k + \mu_2^{k+1}(C1_n - 1_n)$;
        Update $\mu_2^{k+1}$ by $\mu_2^{k+1} = min(\mu_{max}, \rho\mu_2^{k+1})$;
Update the optimal $C$ by Eq.(28);
    repeat
        Update $Z_1^{k+1}$ by $Z_1^{k+1} = Z_1^k + \mu_1^{k+1}(Y - Q - N)$;
        Update $\mu_1^{k+1}$ by $\mu_1^{k+1} = min(\mu_{max}, \rho\mu_1^{k+1})$;
Update the optimal $Q$ by Eq.(31);
Update the optimal $N$ by Eq.(33);
Update the optimal $\alpha_k$ by $\frac{1+\sqrt(4\alpha_{k-1}^2+1)}{2}$;
$k = k + 1$
until Convergence;
Output $W$.
End
```

Table 1: Description of multi-label data sets.

No.s	Data sets	Instances	Training	Test	Features	Labels	Domain
1	Birds	645	322	323	260	19	Audio
2	CAL500	502	251	251	68	174	Music
3	Corel5k	5000	4500	500	499	574	Image
4	Education	5000	2000	3000	550	33	Text
5	Emotions	593	391	202	72	6	Music
6	Genbase	662	463	199	1186	27	Biology
7	Recreation	5000	2000	3000	606	22	Text
8	Scene	2407	1211	1196	294	6	Image
9	Science	5000	2000	3000	743	40	Text
10	Society	5000	2000	3000	636	27	Text
11	Yeast	2417	1499	918	103	14	Biology

4.7. Complexity analysis

In our proposed method, $X \in R^{d\times m}$, $Y \in \{0,1\}^{n\times q}$, and $W \in R^{d\times q}$, where n is the number of instances, d is is the ndimensionality and q is is the number of labels. At each iteration of Algorithm 1, the main computational complexity includes matrix multiplication, matrix inversion and singular value decomposition operations. In step 5, the calculation of Lipshictz constants leads to a complexity of $O\left(qn^2 + n^3\right)$, In step 6, 7 and 8, the cost complexity of updating b and W is $O\left(nd + nq + qd + ndq + nd^2 + n^3\right)$ due to the matrix multiplication and matrix inversion.In step 8, the complexity of updating P is $O\left(qn^2 + dn^2\right)$. Moreover, the computational complexity of updating C, Q and N is $O\left(qn^2 + r_1n^2 + r_2nq\right)$ owing to the matrix multiplication and SVD,, where r_1 and r_2 represent the rank of matrix C and Q respectively.Thus, The total complexity of Algorithm 1 is $O\left(nd + nq + qd + ndq + nd^2 + qn^2 + r_1n^2 + r_2nq + n^3\right)$.

5. Experiments

In the section,we choose five multi-label algorithms for comparative experiments on ten common multi-label datasets so as to demonstrate the effectiveness of the proposed algorithm.

5.1. Data sets

For comprehensive performance evaluation, we use eleven real world multi-label data sets in experiments, including Birds, CAL500, Corel5k, Education, Emotions, Genbase, Recreation, Scene, Science, Society, Yeast. For each data set, the detailed characteristics containing Instances (the number of instances), Training(the number of train instances), Test(the number of test instances), Features (the number of the features), Labels (the number of label classes) and Domain (the domain) are illustrated in Table 1.

5.2. Comparison methods

For demonstrate the preponderance of the proposed method, we implemented seven comparative multi-label leanring methods. Detailed instructions of them are summarized as follows.

MLkNN: A lazy learning approach to multilabel learning.

Glocal: It can simultaneously recover the missing labels, train the linear classifiers, explore and exploit both global and local label correlations.

LLSF: Learning Label-Specific Features for multi-label learning.

LSML:

The configuration parameters of the seven algorithms for comparison are suggested by their original literature.

5.3. Evaluation metrics

To evaluate the performance of multi-label learning, we employ some widely used multi-label evaluation metrics, including AUC, Ranking Loss, Coverage, Average Precision, Hamming Loss, One Error, MacroF1, MicroF1. Given a testing data set $D = \{(x_i, y_i)\}_{i=1}^{n}$, where $y_i \in \{0,1\}_q$ indicates the ground truth labels of ith test instance, and $\overline{y}_i$ is its predicted labels.Let TE_i^+ and TE_i^- be the sets of ground-truth positive and negative labels, and PE_i^+, PE_i^- be the sets of predicted positive and negative labels associated with the ith example associated with the ith example

AUC evaluates the average AUC of all the class labels.

$$\text{AUC} \ = \frac{1}{l}\sum_{i=1}^{1}\frac{\left|\left\{(\mathbf{x}', \mathbf{x}'') \mid f\left(\mathbf{x}', y_j\right) \geq f\left(\mathbf{x}'', y_j\right), (\mathbf{x}', \mathbf{x}'') \in Z_j \times \bar{Z}_j\right\}\right|}{\left|Z_j\right|\left|\bar{Z}_j\right|} \tag{34}$$

where $Z_j = \left\{\mathbf{x}_i \mid y_j \in y_i, 1 \leq i \leq l\right\}\left(\bar{Z}_j = \left\{\mathbf{x}_i \mid y_j \notin Y_i, 1 \leq i \leq l\right\}\right)$ indicates the set of test instances with(without) label y_i.

Ranking Loss evaluates the fraction of reversely ordered label pairs, i.e. an irrelevant label is ranked higher than a relevant label.

$$\text{RankingLoss} = \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{|\text{Set}\, R_i|}{\left|TE_i^+\right|\left|TE_i^-\right|} \tag{35}$$

where $\text{SetR}_i = \left\{(u, v) \mid f_u(\mathbf{x}_i) \leq f_v(\mathbf{x}_i), (u, v) \in TE_i^+ \times TE_i^-\right\}$.

Coverage It refers to the number of steps an algorithm needs to cover all relevant labels of an instance.

Average Precision evaluates the average fraction of relevant labels ranked higher than a particular label $y \in y_i$.

$$\text{Ap} = \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{1}{\left|TE_i^+\right|}\sum_{j \in TE_i^+}\frac{\left|\text{Set}\, P_{ij}\right|}{\text{rank}_F(\mathbf{x}_i, j)} \tag{36}$$

where $\text{SetP}_{ij} = \left\{k \in TE_i^+ \mid \text{rank}_F(\mathbf{x}_i, k) \leq \text{rank}_F(\mathbf{x}_i, j)\right\}$.

Hamming Loss valuates how many times an instance-label pair is misclassified, i.e., a label not belonging to the instance is predicted or a label belonging to the instance is not predicted.

$$\text{HammingLoss} = \frac{1}{n_t l}\sum_{i=1}^{n_t}\sum_{j=1}^{l}[\![h_j(\mathbf{x}_i) \neq y_{ij}\Bigg] \tag{37}$$

One Error evaluates the fraction of instances whose top-ranked label is not in the relevant label set, is an indication function

$$\text{One Error} \ = \frac{1}{n}\sum_{i=1}^{n}[\![\left[\arg\max_{y \in \mathcal{Y}} f(\mathbf{x}_\mathbf{i}, y)\right] \in y_i]\!] \tag{38}$$

MacroF1

MicroF1

For AUC and Average Precision, the higher the value, the better the performance; whereas for the others, the lower the value, the better the performance.

5.4. Experimental results

In the section,we select top ranked attributes as feature substes to prove the performance of each comparable algorithm and the algorithm.Besides,considering that different data has various sizes ,we install 0.1 value for the proportion of the quantity of nearest neighbors in ML-KNN so as to avoid the loss of revelant information.The trade-off parameters are choosed by using "grid search" strategy from $10^{-}310^{-}210^{-}111010^{2} intheexperiment.Asshownintable5-2,5-3,5-4,5-5,theyrespectivelypresentcomparativeresultsofsixalgorit$

Comparison results of multi-label feature selection methods in terms of *Hamming Loss* (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	0.0556	0.0676	0.0674	0.0677	0.0621	0.0557	0.0688
Business	0.0282	0.0287	0.0287	0.0287	0.0285	0.0285	0.0284
Computers	0.0413	0.0435	0.0436	0.0434	0.0430	0.0422	0.0412
Education	0.0421	0.0438	0.0438	0.0438	0.0442	0.0438	0.0417
Emotions	0.2336	0.2402	0.2635	0.2384	0.3284	0.2452	0.2676
Entertainment	0.0613	0.0662	0.0646	0.0667	0.0670	0.0639	0.0650
Health	0.0440	0.0491	0.0487	0.0495	0.0499	0.0480	0.0439
Recreation	0.0611	0.0649	0.0651	0.0648	0.0650	0.0636	0.0650
Reference	0.0301	0.0350	0.0347	0.0350	0.0339	0.0342	0.0302
Scene	0.1208	0.1549	0.1753	0.1606	0.1552	0.1452	0.1238
Science	0.0347	0.0356	0.0356	0.0356	0.0356	0.0354	0.0355
Society	0.0579	0.0603	0.0604	0.0603	0.0601	0.0593	0.0586
Yeast	0.2064	0.2233	0.2226	0.2226	0.2119	0.2122	0.2117
average	0.0797	0.0874	0.0907	0.0877	0.0933	0.0845	0.0847

Comparison results of multi-label feature selection methods in terms of *Ranking Loss* (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	0.1381	0.1421	0.1447	0.1453	0.1547	0.1467	0.1352
Business	0.0442	0.0475	0.0472	0.0479	0.0458	0.0476	0.0463
Computers	0.0981	0.1029	0.1039	0.1028	0.0967	0.1015	0.0994
Education	0.0965	0.1052	0.1044	0.1057	0.1083	0.1065	0.0942
Emotions	0.1908	0.2090	0.2310	0.2048	0.4147	0.2053	0.2563
Entertainment	0.1273	0.1425	0.1407	0.1477	0.1442	0.1344	0.1383
Health	0.0646	0.0775	0.0765	0.0779	0.0780	0.0738	0.0657
Recreation	0.1912	0.2119	0.2159	0.2121	0.2114	0.2057	0.2151
Reference	0.0899	0.1014	0.1022	0.1011	0.0954	0.0989	0.0952
Scene	0.1304	0.2494	0.3058	0.2583	0.2115	0.1836	0.1444
Science	0.1423	0.1543	0.1568	0.1544	0.1537	0.1489	0.1464
Society	0.1526	0.1609	0.1615	0.1608	0.1601	0.1581	0.1531
Yeast	0.1822	0.2023	0.2003	0.1996	0.1853	0.1867	0.1843
average	0.1214	0.1413	0.1479	0.1422	0.1540	0.1327	0.1299

Comparison results of multi-label feature selection methods in terms of *Coverage* (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	3.6900	3.6856	3.7485	3.7427	4.0458	3.8904	3.5343
Business	2.4213	2.5315	2.5379	2.5457	2.5015	2.5465	2.5002
Computers	4.6556	4.8351	4.8650	4.8318	4.5720	4.7916	4.7008
Education	4.0887	4.3828	4.3524	4.4042	4.5027	4.4224	3.9981
Emotions	2.0177	2.1289	2.2365	2.0873	3.0691	2.1152	2.3276
Entertainment	3.3890	3.7111	3.6781	3.8267	3.7453	3.5297	3.6237
Health	3.4416	3.9180	3.8930	3.9405	3.9149	3.7766	3.5095
Recreation	5.0805	5.4935	5.5735	5.5013	5.4738	5.3753	5.5653
Reference	3.4750	3.8531	3.8756	3.8437	3.6573	3.7755	3.6472
Scene	0.7576	1.3505	1.6367	1.3952	1.1655	1.0264	0.8238
Science	7.1232	7.6388	7.7309	7.6447	7.5561	7.4041	7.3058
Society	5.9426	6.1928	6.2199	6.1942	6.1678	6.0986	5.9830
Yeast	6.5610	6.7942	6.6983	6.8061	6.6082	6.6346	6.5941
average	3.9636	4.2519	4.2894	4.2719	4.2922	4.1676	4.0457

Comparison results of multi-label feature selection methods in terms of *Average Precision* (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	0.6764	0.6378	0.6338	0.6306	0.6375	0.6665	0.6274
Business	0.8698	0.8617	0.8625	0.8610	0.8640	0.8619	0.8658
Computers	0.6142	0.6010	0.6001	0.6011	0.6112	0.6057	0.6113
Education	0.5327	0.4924	0.4957	0.4925	0.4802	0.4901	0.5452
Emotions	0.7710	0.7564	0.7332	0.7585	0.6022	0.7562	0.7185
Entertainment	0.5487	0.5049	0.5116	0.4897	0.4969	0.5291	0.5138
Health	0.6742	0.6298	0.6347	0.6265	0.6277	0.6440	0.6763
Recreation	0.4670	0.3946	0.3854	0.3934	0.3923	0.4197	0.3881
Reference	0.6152	0.5748	0.5731	0.5740	0.5915	0.5890	0.5988
Scene	0.7921	0.6593	0.5932	0.6462	0.6874	0.7165	0.7791
Science	0.4450	0.3980	0.3913	0.3975	0.3961	0.4158	0.4303
Society	0.5717	0.5501	0.5485	0.5497	0.5495	0.5589	0.5716
Yeast	0.7463	0.7199	0.7214	0.7211	0.7400	0.7402	0.7417
average	0.6548	0.6155	0.6083	0.6124	0.6070	0.6312	0.6400

Comparison results of multi-label feature selection methods in terms of $Macro-F1$ (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	0.1090	0.0693	0.0802	0.0700	0.0718	0.0909	0.0890
Business	0.1353	0.0997	0.1072	0.0987	0.1121	0.1165	0.0519
Computers	0.0458	0.0224	0.0208	0.0238	0.0521	0.0312	0.0439
Education	0.1044	0.0752	0.0775	0.0733	0.0640	0.0761	0.0573
Emotions	0.5601	0.5477	0.4644	0.5338	0.1412	0.5130	0.4908
Entertainment	0.0998	0.0293	0.0407	0.0139	0.0246	0.0640	0.0641
Health	0.1965	0.1223	0.1331	0.1208	0.1019	0.1263	0.1291
Recreation	0.0776	0.0167	0.0051	0.0179	0.0132	0.0375	0.0073
Reference	0.1120	0.0804	0.0815	0.0794	0.0931	0.0971	0.0399
Scene	0.5878	0.3103	0.1516	0.2826	0.3611	0.4119	0.5778
Science	0.0393	0.0046	0.0026	0.0048	0.0050	0.0201	0.0163
Society	0.0455	0.0179	0.0176	0.0166	0.0172	0.0260	0.0365
Yeast	0.3258	0.2410	0.2339	0.2386	0.3007	0.3014	0.3049
average	0.1968	0.1350	0.1176	0.1297	0.1121	0.1562	0.1585

Comparison results of multi-label feature selection methods in terms of $Micro - F1$ (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	0.4798	0.3439	0.3466	0.3503	0.3699	0.4545	0.3661
Business	0.6781	0.6678	0.6691	0.6675	0.6705	0.6734	0.6757
Computers	0.3559	0.3348	0.3421	0.3367	0.3495	0.3517	0.3597
Education	0.1363	0.0361	0.0396	0.0324	0.0041	0.0397	0.1731
Emotions	0.5937	0.5838	0.5126	0.5782	0.1967	0.5722	0.5178
Entertainment	0.2107	0.0665	0.1086	0.0438	0.0424	0.1282	0.1085
Health	0.4155	0.3434	0.3546	0.3441	0.2775	0.3501	0.4376
Recreation	0.1441	0.0228	0.0060	0.0251	0.0185	0.0666	0.0105
Reference	0.3304	0.2682	0.2814	0.2607	0.2375	0.2584	0.3265
Scene	0.5980	0.3260	0.1657	0.2980	0.3829	0.4472	0.5833
Science	0.1056	0.0158	0.0149	0.0167	0.0074	0.0419	0.0407
Society	0.2422	0.1836	0.2048	0.1641	0.1749	0.1927	0.2150
Yeast	0.6116	0.5631	0.5556	0.5588	0.5932	0.5900	0.5949
average	0.3965	0.3111	0.2996	0.3043	0.2755	0.3417	0.3666

When the missing rate of labels is 30 percent and the top 60 percent of features are selected,the following results can be found observed in terms of the data in tables 5-2,5-3,5-4 and5-5. (1)As for the evaluated AUC and Rankingloss in tables 5-2 and 5-3,the MLML algorithm proposed in this paper has the best performance on 10 datasets (Arts, birds, computer, education, impressions, entertainment, recreation, science, Slashdot, yeast). And the performance of MLML has been significantly improved on more than half of the data sets. (2)About the estimated index Precision, One-error in tables 5-4 and 5-5, the algorithm (MLML) proposed in this paper has the best performance on seven data sets (Arts, computer, education, recreation, science, Slashdot, and yeast). MDDM$_{s}pcandMDDM_{p}rojaresuperiortoMLMLondataset:$
$Birds, Impressionsandentertainment.Atthesametime, MLNBhasbetterperformancethanMLMLonentertainmentwhileitisinferiorto$
$1and5{-}2soastobeintuitive.Thehorizontalaxisdenotesthenumberofselectedfeatures, andtheverticalaxisshowstheperformanceofev$
$e\frac{12N}{K(k+1)}(\sum_{j=1}^{K} r_j^2 - e\frac{K(K+1)^2}{4}Table5 - 7showsthattheFriedmanstatisticsofeveryevaluatedindexwithacriticalvalueat =$
$0.05significantlevel.Eachevaluatedindexclearlyrejectsthezeroassumptionthatallalgorithmshavethesameperformanceaccordingtot$
$7, thatis, thereareapparentdifferencesintheperformanceofeachalgorithm.eeeSummaryoftheFriedmanstatistics$F$_F$
($K = 7,\ N = 13$) and the critical value with different evaluation and the critical value at =0.05 significance level

Metrics	F_F	Critical value ($\alpha = 0.05$)
Hamming loss	14.94	2.4469
Ranking Loss	11.49	
Coverage	13.21	
Average precision	14.36	
Macro-F1	12.65	
Micro-F1	15.94	

Parameter sensitivity analysis In the experiment of the algorithm,four trade-off parameters are involved: $_{123}and_4.Thispaperch$
$offparametersontheperformanceofthealgorithm.Meanwhile, everyparameterhasfivealternativevalues:$ 0.010.1110100.$Theexper$
3.$Weobviouslyreflecthowtheperformanceofthisalgorithm(MLML)changeswiththealteringofparametersbymeansoffourevaluated$
$3, _1 = 100_2 = 0.1_3 = 0.1and_4 = 0.1, MLMLachievesthebestperformanceonthedatasetYeast.$

6. Conclusion

The approaches of discernibility matrix provide a mathematical foundation for fuzzy rough set-based data analysis. However, a way how to effectively realize the superiority for dealing with multi-label data sets,