Interactive Causal Correlation Space Reshape for Multi-Label Classification

Chao Zhang[1], Yusheng Cheng[1,2], Yibin Wang[1,2], Yuting Xu[1] *

[1] School of Computer and Information, Anqing Normal University, Anhui (China)
[2] The University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing (China)

Received 23 December 2021 | Accepted 2 June 2022 | Published 10 August 2022

Abstract

Most existing multi-label classification models focus on distance metrics and feature spare strategies to extract specific features of labels. Those models use the cosine similarity to construct the label correlation matrix to constraint solution space, and then mine the latent semantic information of the label space. However, the label correlation matrix is usually directly added to the model, which ignores the interactive causality of the correlation between the labels. Considering the label-specific features based on the distance method merely may have the problem of distance measurement failure in the high-dimensional space, while based on the sparse weight matrix method may cause the problem that parameter is dependent on manual selection. Eventually, this leads to poor classifier performance. In addition, it is considered that logical labels cannot describe the importance of different labels and cannot fully express semantic information. Based on these, we propose an Interactive Causal Correlation Space Reshape for Multi-Label Classification (CCSRMC) algorithm. Firstly, the algorithm constructs the label propagation matrix using characteristic that similar instances can be linearly represented by each other. Secondly, label co-occurrence matrix is constructed by combining the conditional probability test method, which is based on the label propagation reshaping the label space to rich label semantics. Then the label co-occurrence matrix combines with the label correlation matrix to construct the label interactive causal correlation matrix to perform multi-label classification learning on the obtained numerical label matrix. Finally, the algorithm in this paper is compared with multiple advanced algorithms on multiple benchmark multi-label datasets. The results show that considering the interactive causal label correlation can reduce the redundant information in the model and improve the performance of the multi-label classifier.

Keywords

Conditional Probability, Interactive Causal Inference, Label Co-Occurrence, Label Space Reshape, Multi-Label Classification.

DOI: 10.9781/ijimai.2022.08.007

I. Introduction

With the continuous development of machine learning, classification models have evolved rapidly. However, in actual scenarios, there are still problems such as unbalanced classification, multi-layer and multi-label. In multi-label classification, labels with complex dependencies are more likely to identify the same instance. Therefore, labels correlation is mostly considered when performing multi-label classification. The consideration of the unique characteristics of labels effectively reduces the difficulty of capturing important information from high-dimensional data to construct competitive classifiers. These all require in-depth exploration of potential associations or dependencies between labels [1].

In order to mine the potential information of the label space, one method is to use the feature sparsity strategy to extract the specific features of the label. Algorithms based on this strategy usually use the l_1-norm to constraint weight matrix. The l_1-norm has high sparseness, so that only some important features contribute to the model, thereby extracting label-specific features. Typical algorithms include the LLSF algorithm (Learning Label-Specific Features for Multi-Label Classification) proposed by Huang et al. [2]. This algorithm assumes that each label is only related to certain features in the original feature space. The highly correlated labels have similarities. The label correlation and l_1-norm sparsity constraints are used to extract label-specific features. Based on the same strategy, the LSF-CI (Multi-Label Learning with Label-Specific Features Using Correlation Information) algorithm proposed by Han et al. [3] assumes that labels are only related to specific features. Features contribute differently to different targets. Similar labels have similar features. The sparse feature weight matrix is constructed by considering the correlation of the targets to extract the label-specific features. The other method is based on a distance measurement strategy. The main idea of this strategy is to find a set of measurement reference points. Then calculate the distance from each label of each sample to these measurement reference points. Finally, the measurement result is used as the label-specific features of each label. The typical algorithm of the measurement strategy is the LIFT algorithm (Multi-Label Learning with Label-Specific Features) proposed by Zhang [4]. The algorithm uses k-Means clustering for

* Corresponding author.
E-mail addresses: 767991942@qq.com (C. Zhang), chengyshaq@163.com (Y. Cheng), wangyb07@mail.ustc.edu.cn (Y. Wang), 1619288826@qq.com (Y. Xu).

positive samples and negative samples respectively. The obtained cluster centers are regarded as the center of the sample. Euclidean distance is used to measure the distances of all samples to these cluster centers to achieve the extraction of label-specific features.

However, some label-specific features extraction algorithms represented by LIFT do not consider label correlation. As an important means of mining features and label latent information, label correlation is introduced into the classification model to effectively improve the accuracy of multi-label classification. Based on this concept, the LSML algorithm (Improving Multi-Label Classification with Missing Labels by Learning Label-Specific Features) proposed by Huang et al. [5] deals with the multi-label classification of the missing dataset by learning high-order label correlation matrix and label-specific features. By learning high-level label correlation, a new supplementary label matrix is augmented from the incomplete label matrix. Then, it learns a label-specific data representation for each type of label. On this basis, combine the learned high-order label correlation to construct a multi-label classifier. The FF-MLLA algorithm (Multi-Label Lazy Learning Approach based on FireFly Method) proposed by Cheng et al. [6] uses the Firefly method to fuse correlation information with sample similarity information. Finally, the classification by singular value decomposition and extreme learning machine has achieved certain results.

However, the existing classification algorithms only consider the degree of correlation between targets when contacting the correlation of targets, but ignore the causality that exists in the interactive causal correlation. Only considering the correlation of symmetric labels will often cause the problem of redundant information [7] in the model, resulting in a decrease in the performance of the classifier. In real life, this asymmetric correlation is also very common. A classic explanation is shown in Fig. 1: The crowing of a rooster symbolizes the coming of dawn. The reason for the rooster's crowing is that the hormones in the rooster's brain are sensitive to light. Therefore, the disappearance of darkness makes this physiological phenomenon occur in roosters. This phenomenon indicates that there is a correlation between dawn and rooster crowing. The dawn is the cause, and the rooster's crowing is the result. However, dawn did not appear because of the rooster's crowing. This phenomenon indicates that the interactive causal correlation between the two is asymmetric.

Fig. 1. Interactive Causal Correlation.

In multi-label learning, label objects with complex dependencies also have similar causality problems. Based on this fact, the ACML algorithm (Asymmetry Label Correlation for Multi-Label Learning) proposed by Bao et al. [8] uses cosine similarity to construct a label correlation matrix, and then measures the adjacency between labels to construct a label adjacency matrix. The label adjacency matrix is used to constrain the label correlation matrix to link the asymmetric label correlation. Considering the case of multi-label learning, the correlation between labels may come from the dependence of labels on the same set of features, or the dependence of one label on another label. We abstract the complex relationship between labels as a relationship--"Co-occurrence" [9]. When two label objects often appear together, we think that they have complex dependencies. The judgment of the degree of dependence between two related variables requires further consideration of directionality. The direction of greater dependence is used as the standard for inferring the interactive causal relationship between labels. Generally, we measure the co-occurrence relationship by the method of conditional probability (conditional independence test) [10]. That is, the probability P (Label2 | Label1), which represents the probability of the appearance of Label2 under the condition of the appearance of Label1. When this probability is greater than a specific threshold, we think that Label1 and Label2 have a dependency relationship. At the same time, the label-specific features take into account the unique characteristics of the label, which alleviates the high-dimensionality problem in multi-label learning to a certain extent. However, the failure of the distance measurement [11] and the problem of the l_1-norm feature sparsity parameter [12] are selected manually still exists. The constraint-based conditional independence test method considers the causal relationship between the labels while further avoiding the problem of Euclidean distance failure in the high-dimensional space. The use of naturally existing dependencies to extract label-specific features can also avoid the problem of l_1-norm relying on manual parameter selection to a certain extent. Based on this, this paper proposes the ***Interactive Causal Correlation Space Reshape for Multi-label Classification (CCSRMC)*** algorithm. By using space reshaping [13] methods to solve logical targets, there are problems such as the inability to describe the importance of different targets and the inability to fully express latent information. On this basis, the label co-occurrence matrix is constructed by combining the conditional probability test method. Then it is combined with the label correlation matrix to construct the label interactive causal correlation matrix to perform multi-label classification learning on the obtained numerical label matrix. Taking into account the sparseness of the label space, the label space reshaping method is used to transform the original discrete label into continuous label, which is used to infer the interactive causal relationship between features and labels. Then it obtains the numerical label matrix and extracts the label-specific features for multi-label classification. The algorithm in this paper carries out comparative experiments and statistical hypothesis tests with multiple advanced algorithms on multiple benchmark multi-label data sets. It also conducts ablation analysis with or without consideration of causality. The results of experiments and analysis verify the effectiveness of our algorithm.

The rest of this paper is organized as follows. In section Ⅱ, we introduce the model of the CCSRMC and the method to complete model optimization. In section Ⅲ, the algorithm pseudocode and complexity analysis of the proposed algorithm is given. In section Ⅳ, datasets, evaluation metrics, parameters setup are introduced. In section Ⅴ, we validate the proposed method with a hypothesis test and the sensitivity of the parameters was analyzed. Meanwhile experimental results on 14 benchmark datasets are given. Finally, we conclude our work in section Ⅵ.

II. The Proposed Method

A. Interactive Causal Inference Theory

Many scholars believe that there is a certain dependency in the correlation between things, which may lead to the asymmetry of the correlation. Scholars collectively call it causal inference [14]. Early causal inference algorithms can usually be divided into two categories: Constraint Based method and Function Based method. The constraint-based method is also called the method based on the independence test. The basic idea is to transform the inference problem of the direction of dependence into the problem of judging the degree of dependence between two variables. This type of algorithm first calculates the dependence of the two directions at the same time, and then takes the direction with the greater dependence as the inferred interactive causal direction. Common algorithms based on conditional independent tests include Granger Causality (GC) [15], which is a classic causality

discovery tool, but it is only applicable to Gaussian cases. TE (Transfer Entropy) [16] is a non-linear promotion of GC. It uses the concept of information theory, which is equivalent to Conditional Mutual Information (CMI) [17]. In addition, you can also use the Kernel function [18] method and distance correlation [19] perform conditional independent testing. K2 search algorithm (K2 Search), PC algorithm (Peter-Clark) and IC algorithm (Inductive Causation) [20], etc. Causal inference algorithms usually have high algorithm complexity and poor adaptability to high-dimensional data in multi-label learning. In this paper, an interactive causal inference method based on conditional independence tests is used to infer the potential dependencies between multi-label learning. The label interactive causal inference method is a constraint-based algorithm, which ignores the influence of Confounder Variables. However, because only the dependency between the pair of variables is considered, the algorithm has low complexity and fast calculation speed, and can better handle high-dimensional data in multi-label learning.

Herein we introduce the label interactive causal inference method based on conditional independence test:

For a pair of variables (a, b). P(a=n) represents the probability when a=n. P(b|a=n) represents the conditional probability of variable b when a=n, let λ=P(a), μ=P(b|a). The label interactive causality inference method treats (λ, μ) as two independent random variables. The distance correlation coefficients of the two possible directions are calculated separately. The direction with the smaller coefficient is used as the inferred interactive causal direction.

Let f_λ and f_μ respectively denote the characteristic function of (λ, μ). $f_{\lambda\mu}$ is (λ, μ) joint characteristic function, then the distance covariance $\mathcal{C}^2(\lambda, \mu)$ of (λ, μ) is:

$$\mathcal{C}^2(\lambda,\mu) = \left\| f_{\lambda,\mu} - f_\lambda f_\mu \right\|^2 \tag{1}$$

The distance correlation coefficient $\mathcal{D}(\lambda, \mu)$ is:

$$\mathcal{D}(\lambda,\mu) = \frac{\mathcal{C}(\lambda,\mu)}{\sqrt{\mathcal{C}(\lambda,\lambda)\mathcal{C}(\mu,\mu)}} \tag{2}$$

If $\mathcal{C}(\lambda,\lambda)=0$ or $\mathcal{C}(\mu,\mu)=0$, then $\mathcal{D}(\lambda,\mu)=0$. Suppose the multi-label dataset contains n instances, and l labels, then for any pair of labels $(X_i, Y_j \mid i, j = 1, 2, 3, \ldots, l)$, n groups of variables $\{(\lambda_i, \mu_j)\}_{i,j=1}^{n}$ can be constructed. For variables λ and μ, matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are constructed as follows:

$$\boldsymbol{A}_{ij} = \left\|\lambda_i - \lambda_j\right\| - \frac{1}{n}\sum_{j=1}^{n}\lambda_{ij} - \frac{1}{n}\sum_{i=1}^{n}\lambda_{ij} + \frac{1}{n^2}\sum_{i,j=1}^{n}\lambda_{ij} \tag{3}$$

$$\boldsymbol{B}_{ij} = \left\|\mu_i - \mu_j\right\| - \frac{1}{n}\sum_{j=1}^{n}\mu_{ij} - \frac{1}{n}\sum_{i=1}^{n}\mu_{ij} + \frac{1}{n^2}\sum_{i,j=1}^{n}\mu_{ij} \tag{4}$$

The distance covariance can be calculated by:

$$\mathcal{C}_n(\lambda,\mu) = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n}\boldsymbol{A}_{ij}\boldsymbol{B}_{ij}} \tag{5}$$

From formula (2) and formula (5), we can know:

$$\mathcal{D}(\lambda,\mu) = n\frac{\sqrt{\sum_{i,j=1}^{n}\boldsymbol{A}_{ij}\boldsymbol{B}_{ij}}}{\sqrt{\sum_{i,j=1}^{n}\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}}\sqrt{\sum_{i,j=1}^{n}\boldsymbol{B}_{ij}\boldsymbol{B}_{ij}}} \tag{6}$$

$\mathcal{D}(\lambda, \mu)$ is always greater than 0. For any pair of variables (a, b):

If $\mathcal{D}_{b\to a} > \mathcal{D}_{a\to b}$, then $a \to b$ is the inferred interactive causal direction

If $\mathcal{D}_{a\to b} > \mathcal{D}_{b\to a}$, then $b \to a$ is the inferred interactive causal direction

B. Establishment of the CCSRMC Model

In multi-label learning [21], there are input training data $\boldsymbol{X}$ and label matrix $\boldsymbol{Y}$, where $\boldsymbol{X}\in\mathbb{R}^{n\times d}$, $\boldsymbol{Y}\in\mathbb{R}^{n\times l}$, l is the number of labels, n is the number of training samples, d is the number of features. $\boldsymbol{U}=\{(x_i, y_i) \mid 1\le i\le n\}$ is a multi-label training dataset, where $x_i = \{\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^d\}$ is the i-th feature vector, $y_i = \{\boldsymbol{y}_i^1, \boldsymbol{y}_i^2, \cdots, \boldsymbol{y}_i^l\}$ is the i-th label vector. The task of multi-label learning is to find a mapping relationship $f: \boldsymbol{X}\to 2^Y$. The general multi-label algorithm [22] model is:

$$\boldsymbol{L}(\boldsymbol{W}) = \min_{\boldsymbol{W}} \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_F^2 + \beta\|\boldsymbol{W}\|_1 \tag{7}$$

$\boldsymbol{L}(\boldsymbol{W})$ is the loss function, β is the regularization parameter, and $\boldsymbol{W}\in\mathbb{R}^{d\times l}$ is the weight matrix.

Based on the original multi-label algorithm model, the algorithm in this paper uses the feature of linear representation between similar instances to construct the label propagation matrix $\boldsymbol{P}$. We use formula (8) to calculate an $\boldsymbol{N}\times\boldsymbol{N}$ similarity matrix $\boldsymbol{A}$ between $\boldsymbol{N}$ instances:

$$\boldsymbol{A}_{jk} = \begin{cases} \boldsymbol{exp}\left(-\dfrac{\left\|x_j - x_k\right\|^2}{2}\right) & \text{if } j \ne k \\ 0 & \text{if } j = k \end{cases} \tag{8}$$

Where $\boldsymbol{D} = \mathrm{diag}[\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots \boldsymbol{d}_m]$, $\boldsymbol{d}_j = \sum_{k=1}^{m}\boldsymbol{W}_{jk}$.

$$\boldsymbol{P} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}} \tag{9}$$

The original label matrix $\boldsymbol{Y}$ is projected into a new numerical label matrix $\boldsymbol{S}$ through the label propagation matrix $\boldsymbol{P}$ to perform multi-label classification learning. After the introduction of label correlation optimization, formula (7) can be expressed as:

$$\min_{W_i}\frac{1}{2}\|\boldsymbol{X}\boldsymbol{W}_i - \boldsymbol{S}\|_2^2 + \frac{1}{2}\|\boldsymbol{Y}\boldsymbol{P} - \boldsymbol{S}\|_F^2 + \frac{\alpha}{2}\sum_{j=1}^{l}\boldsymbol{R}_{ij}\boldsymbol{W}_i^T\boldsymbol{W}_j + \beta\|\boldsymbol{W}\|_1 \tag{10}$$

The first term of equation (10) is to minimize the error of the sum of squares. The numerical label vector $\boldsymbol{S}_i$ is used instead of the logical label vector $\boldsymbol{Y}_i$. Numerical labels bring more semantic information and are more conducive to the correlation between contact labels. The second term completes the label propagation to the original label matrix $\boldsymbol{Y}$. The third term indicates that the strong correlation between the label y_i and the label y_j leads to a great similarity between $\boldsymbol{W}_i$ and $\boldsymbol{W}_j$, where $\boldsymbol{R}\in\mathbb{R}^{l\times l}$ is the label correlation matrix, which is calculated by cosine similarity. However, some existing multi-label learning algorithms usually directly add the label correlation matrix to the model to constrain the solution space when considering the label correlation. These algorithms ignore the asymmetry of the correlation relationship between labels. The algorithm in this paper uses the label correlation matrix and the label co-occurrence matrix to construct the label interactive causal correlation matrix. The label co-occurrence matrix analyzes the potential interactive causal relationship between the labels through the conditional probability test method. Considering all binary classifiers at the same time, the optimization can be expressed as:

$$\min_{W}\frac{1}{2}\|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{S}\|_F^2 + \frac{1}{2}\|\boldsymbol{Y}\boldsymbol{P} - \boldsymbol{S}\|_F^2 + \frac{\alpha}{2}tr(\boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^T) + \beta\|\boldsymbol{W}\|_1 \tag{11}$$

Based on the interactive causal inference theory, we get the following two definitions:

Definition 1: $\boldsymbol{V}$ is the label interactive causality correlation matrix, and $\boldsymbol{V}$ is a square matrix with a dimension of $l\times l$. If $\boldsymbol{V}_{mn} = \boldsymbol{V}_{nm} \ne 0$, there is a correlation between the corresponding labels. If $\boldsymbol{V}_{mn} = \boldsymbol{V}_{nm} = 0$, there is neither correlation nor interactive causality between the corresponding labels. If $\boldsymbol{V}_{mn} \ne \boldsymbol{V}_{nm}$ and $\boldsymbol{V}_{mn}$ or $\boldsymbol{V}_{nm} = 0$, it means that there is an interactive causal relationship between the corresponding labels.

Definition 2: $\boldsymbol{R} \in \mathbb{R}^{l\times l}$ is the label correlation matrix. $\boldsymbol{C} \in \{0,1\}^{l\times l}$ is the label co-occurrence matrix. $\boldsymbol{A} \odot \boldsymbol{B}$ is defined as a matrix operation, where $\boldsymbol{A} \in \mathbb{R}^{l\times l}$, $\boldsymbol{B} \in \{0,1\}^{l\times l}$. $\odot$ means that if $\boldsymbol{B}_{mn} = 1$, then $\boldsymbol{A}_{mn} = 0$. $\boldsymbol{C}^C$ is the complement matrix of the label co-occurrence matrix $\boldsymbol{C}$. Then the label interactive causality correlation matrix $\boldsymbol{V}$ is:

$$\boldsymbol{V} = \boldsymbol{R} \odot \boldsymbol{C}^C \tag{12}$$

y_m, y_n are a pair of labels, where $i, j = 1, 2, ..., n$. According to the description of label interactive causal inference in section Ⅱ-A. Let $\lambda = \mathrm{P}(y_m)$, $\mu = \mathrm{P}(y_n | y_m)$. Assuming (λ, μ) is a pair of independent random variables, f_λ and f_μ are corresponding to its characteristic function, $f_{\lambda,\mu}$ is its joint characteristic function. According to section Ⅱ-A:

If $\mathcal{D}_{y_n \to y_m} > \mathcal{D}_{y_m \to y_n}$, then $y_m \to y_n$ is the inferred interactive causal direction.

If $\mathcal{D}_{y_m \to y_n} > \mathcal{D}_{y_n \to y_m}$, then $y_n \to y_m$ is the inferred interactive causal direction.

In summary, we can get the label interactive causality correlation matrix $\boldsymbol{V} \in \{0,1\}^{l\times l}$, where $\boldsymbol{V}_{mn} = 0$ means that there is no interactive causal relationship between the m-th label and the n-th label. $\boldsymbol{V}_{mn} = 1$ means that the direction of inferring the interactive causal relationship between the m-th label and the n-th label is $m \to n$. Finally, the obtained label interactive causality correlation matrix is added to formula (11) to obtain the algorithm model proposed in this chapter:

$$\min_{W} \frac{1}{2}\|\boldsymbol{XW} - \boldsymbol{S}\|_F^2 + \frac{1}{2}\|\boldsymbol{YP} - \boldsymbol{S}\|_F^2 + \frac{\alpha}{2} tr(\boldsymbol{WVW}^T) + \beta\|\boldsymbol{W}\|_1 \tag{13}$$

Where $\boldsymbol{W} = (\boldsymbol{W}_1, \boldsymbol{W}_2, ..., \boldsymbol{W}_l) \in \boldsymbol{R}^{d\times l}$, $\boldsymbol{S} = (\boldsymbol{S}_1, \boldsymbol{S}_2, ..., \boldsymbol{S}_l) \in \boldsymbol{R}^{n\times l}$, $\alpha > 0$ and $\beta > 0$ are both parameters in the algorithms model. Then constraints to the reshaping process are considered to add, so that the reshaped label matrix $\boldsymbol{S}$ after the mapping has a small difference from the original label matrix $\boldsymbol{Y}$:

$$\begin{aligned}\min_{W} &\frac{1}{2}\|\boldsymbol{XW} - \boldsymbol{S}\|_F^2 + \frac{1}{2}\|\boldsymbol{YP} - \boldsymbol{S}\|_F^2 + \frac{\alpha}{2} tr(\boldsymbol{WVW}^T) \\ &+ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{S}\|_F^2 + \beta\|\boldsymbol{W}\|_1\end{aligned} \tag{14}$$

Using natural interactive causality to extract label-specific features can avoid the problem that the $\boldsymbol{l}_1$-norm feature sparse strategy relies on manual parameter selection to a certain extent. Too high or too low sparsity will lead to poor classifier performance. In addition, in order to prevent the algorithm from overfitting that may be caused by the reshaping of the numerical label matrix $\boldsymbol{S}$. This paper uses F-norm to constrain matrix $\boldsymbol{S}$ and control the sparsity of matrix $\boldsymbol{W}$. In summary, the CCSRMC algorithm model proposed in this paper is as follows:

$$\begin{aligned}\min_{W} &\frac{1}{2}\|\boldsymbol{XW} - \boldsymbol{S}\|_F^2 + \frac{\alpha}{2} tr(\boldsymbol{WVW}^T) + \frac{1}{2}\|\boldsymbol{YP} - \boldsymbol{S}\|_F^2 \\ &+ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{S}\|_F^2 + \beta||\boldsymbol{W}||_F^2 + \gamma\|\boldsymbol{P}\|_{2,1} + \frac{\alpha}{2}\|\boldsymbol{S}\|_F^2\end{aligned} \tag{15}$$

Where $\gamma > 0$ is the parameter in the algorithm model. It can be seen from the above algorithm model that when the label matrix is reshaped, it will be affected by the weight $\boldsymbol{W}$. What affects the weight $\boldsymbol{W}$ is not only the classification model after the label matrix is reshaped, but also the correlation between the labels [23]. The label correlation matrix is usually directly added to the model without considering the interactive causality of the correlation between the labels. As shown in Fig. 2, we construct a label co-occurrence matrix by combining conditional probability testing method on the basis of label propagation reshaping the rich label semantics in label space. The label co-occurrence matrix and the label correlation matrix are combined to construct the numerical label matrix obtained by the label interactive causal correlation matrix pair, so that multi-label classification learning is performed.

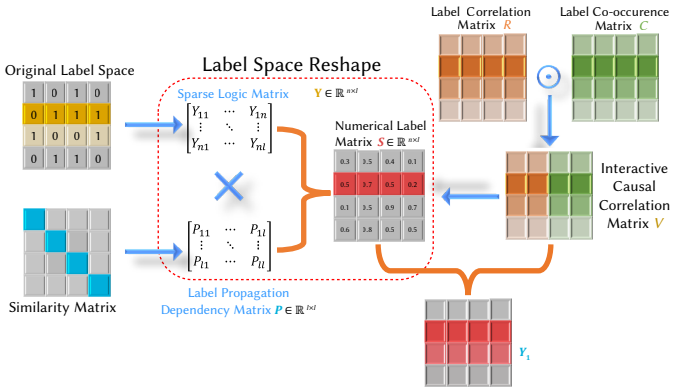

Fig. 2. Interactive Causal Correlation Space Reshape.

C. Optimization of the Model

In this paper, the three variable matrices in the label space reshaping model are calculated by using the alternate iteration method. That is the label propagation dependency matrix $\boldsymbol{P}$, the numerical labels matrix $\boldsymbol{S}$, and the weight matrix $\boldsymbol{W}$ are used to complete the optimization of the entire model and refer to the literature [24].

$$\min_{P} \frac{1}{2}\|\boldsymbol{YP} - \boldsymbol{S}\|_F^2 + \gamma\|\boldsymbol{P}\|_{2,1} \tag{16}$$

Where γ is parameter of the model. From the above algorithm model, it can be seen the numerical label matrix $\boldsymbol{S}$ is constantly changing. This matrix has correlation between labels, which means that when learning label-specific features, it will take the relevance between different labels into account. For the label propagation dependency matrix $\boldsymbol{P}$, we can define $||\boldsymbol{P}||_{2,1} = \sum_{j=1}^{l}\sqrt{\sum_{i=1}^{n}(\boldsymbol{P}_{ij})^2}$, Through the label matrix $\boldsymbol{S}$ reshaped by the $\boldsymbol{l}_{2,1}$-norm sparse label space of the matrix, the variable matrix $\boldsymbol{P}$ is obtained as:

$$\nabla\boldsymbol{P} = \boldsymbol{Y}^T\boldsymbol{YP} - \boldsymbol{Y}^T\boldsymbol{S} + 2\gamma\boldsymbol{DP} = 0 \tag{17}$$

$$\boldsymbol{P} = (\boldsymbol{Y}^T\boldsymbol{Y} + 2\gamma\boldsymbol{D})^{-1}\boldsymbol{Y}^T\boldsymbol{S} \tag{18}$$

Where $\boldsymbol{D}_{ii} = \frac{1}{2\|\boldsymbol{P}_{i*}\|_2}$, $\boldsymbol{D}$ is the diagonal matrix. Then calculate the numerical label matrix $\boldsymbol{S}$, the objective function is:

$$\min_{S} \frac{1}{2}\|\boldsymbol{XW} - \boldsymbol{S}\|_F^2 + \frac{1}{2}\|\boldsymbol{YP} - \boldsymbol{S}\|_F^2 + \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{S}\| + \frac{\alpha}{2}\|\boldsymbol{S}\|_F^2 \tag{19}$$

In training, in order to minimize the risk of label reshaping model structure and prevent the occurrence of overfitting, the model parameter α appears in the objective function as the control parameter of the model weight penalty term. The number label matrix $\boldsymbol{S}$ is obtained as:

$$\boldsymbol{S} = \frac{1}{\alpha + 3}(\boldsymbol{XW} + \boldsymbol{YP} + \boldsymbol{Y}) \tag{20}$$

Finally, the weight matrix $\boldsymbol{W}$ is calculated. The objective function is:

$$\min_{W} \frac{1}{2}\|\boldsymbol{XW} - \boldsymbol{S}\|_F^2 + \frac{\alpha}{2} tr(\boldsymbol{W}^T\boldsymbol{VW}) + \beta\|\boldsymbol{W}\|_1 \tag{21}$$

The composite function derivation is performed on the weight matrix $\boldsymbol{W}$. Split $\boldsymbol{W}$ into $f(\boldsymbol{W})$ and $g(\boldsymbol{W})$:

$$f(\boldsymbol{W}) = \min_{W} \frac{1}{2}\|\boldsymbol{XW} - \boldsymbol{S}\|_F^2 + \frac{\alpha}{2} tr(\boldsymbol{W}^T\boldsymbol{VW}) \tag{22}$$

$$g(\boldsymbol{W}) = \beta\|\boldsymbol{W}\|_1 \tag{23}$$

When learning tag information, the correlation between labels should be considered. The third item in the weight matrix $\boldsymbol{W}$ model is to use the F-norm to sparse the weight matrix $\boldsymbol{W}$. The parameter

β controls the sparsity of the weight matrix. Although formula (21) is a convex optimization problem, the objective function (21) is non-smooth due to the non-smoothness of the regular term of $\boldsymbol{l}_1$-norm. For this reason, this paper uses the near-end gradient descent method [22] to solve the optimization problem of non-smooth objective function. The objective function becomes:

$$\min_{W\in\mathcal{H}} F(\boldsymbol{W}) = f(\boldsymbol{W}) + g(\boldsymbol{W}) \tag{24}$$

$$\nabla f(\boldsymbol{W}) = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W} - \boldsymbol{X}^T\boldsymbol{S} + \alpha\boldsymbol{W}\boldsymbol{V} \tag{25}$$

In formula (24), $\mathcal{H}$ is the Hilbert space. Both $f(\boldsymbol{W})$ and $g(\boldsymbol{W})$ are convex functions and satisfy Lipschitz condition. For any matrix $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ there are:

$$\|\nabla f(\boldsymbol{W}_1) - \nabla f(\boldsymbol{W}_2)\| \le L_g\|\Delta\boldsymbol{W}\| \tag{26}$$

Where L_g is Lipschitz constant. $\Delta\boldsymbol{W} = \boldsymbol{W}_1 - \boldsymbol{W}_2$. In the process of accelerating gradient descent, $F(\boldsymbol{W})$ is no longer directly minimized. It is necessary to introduce the quadratic approximation $F(\boldsymbol{W})$ of $\boldsymbol{Q}(\boldsymbol{W}, \boldsymbol{W}^{(t)})$, so define $\boldsymbol{Q}(\boldsymbol{W}, \boldsymbol{W}^{(t)})$:

$$Q\left(\boldsymbol{W}, \boldsymbol{W}^{(t)}\right) = f\left(\boldsymbol{W}^{(t)}\right) + \left(\nabla f\left(\boldsymbol{W}^{(t)}\right), \boldsymbol{W} - \boldsymbol{W}^{(t)}\right) + \frac{L_g}{2}\left\|\boldsymbol{W} - \boldsymbol{W}^{(t)}\right\|_F^2 + g(\boldsymbol{W}) \tag{27}$$

$$\boldsymbol{G}_t(\boldsymbol{W}) = \boldsymbol{W}_t - \frac{1}{L_g}\nabla f(w) \tag{28}$$

$$\begin{aligned}\boldsymbol{W} &= \arg\min_{W} \boldsymbol{Q}\left(\boldsymbol{W}, \boldsymbol{W}^{(t)}\right)\\ &= \arg\min_{W} g(\boldsymbol{W}) + \frac{L_g}{2}\left\|\boldsymbol{W} - \boldsymbol{G}^{(t)}\right\|_F^2\\ &= \arg\min_{W} \frac{1}{2}\left\|\boldsymbol{W} - \boldsymbol{G}^{(t)}\right\|_F^2 + \frac{\alpha}{L_g}\|\boldsymbol{W}\|_1\end{aligned} \tag{29}$$

Given in the research in [25]:

$$\boldsymbol{W}^{(t)} = \boldsymbol{W}_t + \frac{b_{t+1} - 1}{b_t}(\boldsymbol{W}_t - \boldsymbol{W}_{t-1}) \tag{30}$$

The sequence b_t satisfies $b_{t+1}^2 - b_{t+1} \le b_t^2$, which can increase the speed of convergence to $\boldsymbol{O}(t^{-2})$. So $\boldsymbol{W}_t$ can be regarded as the result of the t-th iteration of $\boldsymbol{W}$:

$$\boldsymbol{W}_{t+1} = \boldsymbol{S}_\varepsilon\left[\boldsymbol{G}^{(t)}\right] = \arg\min_{W} \varepsilon\|\boldsymbol{W}\|_1 + \frac{1}{2}\left\|\boldsymbol{W} - \boldsymbol{G}^{(t)}\right\|_F^2 \tag{31}$$

Where $\boldsymbol{S}_\varepsilon[\cdot]$ is a soft threshold operator. For any parameter x_{ij} and $\varepsilon = \frac{\alpha}{L_g}$, this function is defined as:

$$\boldsymbol{S}_\varepsilon(x_{ij}) = \begin{cases} x_{ij} - \varepsilon & when\ x_{ij} > \varepsilon \\ x_{ij} + \varepsilon & when\ x_{ij} < -\varepsilon \\ 0 \end{cases} \tag{32}$$

$$\boldsymbol{W} = soft\left(\boldsymbol{G}_t(\boldsymbol{W}), \frac{\beta}{L_g}\right) \tag{33}$$

The Lipschitz constant is calculated as follows. For given $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, the Lipschitz condition is satisfied according to $f(\boldsymbol{W})$:

$$\begin{aligned}\|\nabla f(\boldsymbol{W}_1) - \nabla f(\boldsymbol{W}_2)\|_F^2 &= \left\| \begin{matrix} \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}_1 - \boldsymbol{X}^T\boldsymbol{S} + \alpha\boldsymbol{W}\boldsymbol{V} \\ -(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}_2 - \boldsymbol{X}^T\boldsymbol{S} + \alpha\boldsymbol{W}_2\boldsymbol{R}) \end{matrix} \right\|_F^2\\ &= \|\boldsymbol{X}^T\boldsymbol{X}\Delta\boldsymbol{W} + \alpha\Delta\boldsymbol{W}\boldsymbol{V}\|_F^2\\ &\le \|\boldsymbol{X}^T\boldsymbol{X}\Delta\boldsymbol{W}\|_F^2 + \|\alpha\Delta\boldsymbol{W}\boldsymbol{V}\|_F^2\\ &\le 2\|\boldsymbol{X}^T\boldsymbol{X}\|_2^2\|\Delta\boldsymbol{W}\|_F^2 + 2\alpha\|\boldsymbol{V}\|_2^2\|\Delta\boldsymbol{W}\|_F^2\\ &= (2\|\boldsymbol{X}^T\boldsymbol{X}\|_2^2 + 2\alpha\|\boldsymbol{R}\boldsymbol{V}\|_2^2)\|\Delta\boldsymbol{W}\|_F^2\end{aligned} \tag{34}$$

Therefore, the Lipschitz constant of the model is:

$$L_g = \sqrt{2\|\boldsymbol{X}^T\boldsymbol{X}\|_2^2 + 2\alpha\|\boldsymbol{V}\|_2^2} \tag{35}$$

III. Pseudocode and Complexity Analysis

A. Accelerated Gradient Descent

This section outlines the algorithm flow of CCSRMC. Solve the weight matrix $\boldsymbol{W}$ and obtain the interactive causal correlation matrix $\boldsymbol{V}$. In Algorithm 1 Step 3 and Step 4 are more complicated. Where $\boldsymbol{G}^{(t)}$ is an intermediate variable, $f(\cdot)$ represents the gradient, the algorithm complexity of calculating the weight matrix $\boldsymbol{W}$ is $\boldsymbol{O}(n^2\ d^2\ l + n^2\ dl + dl^2)$. In algorithm 2, the label interactive causality matrix: $\boldsymbol{V}$ is constructed with the conditional independence test method. Only the non-diagonal elements in the upper or lower triangular matrix need to be calculated. Therefore, the complexity of step 3 is $\boldsymbol{O}(l^2)$. Step 4 has a complexity of $\boldsymbol{O}(l^2/2)$.

The accelerated proximal gradient of CCSRMC is summarized in Algorithm 1.

Algorithm 1: The Accelerated Proximal Gradient Method

Input: Training data matrix: $\boldsymbol{X}$; Training label data set: $\boldsymbol{Y}$; Parameters: α, β, γ

Output: Weight matrix: $\boldsymbol{W}$.

1. Initialize: $b_0 = b_1 = 1, \boldsymbol{W}_0 = \boldsymbol{W}_1 = (\boldsymbol{X}^T\boldsymbol{X} + \gamma I)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$
2. **while** not converged **do**
3. $G^{(t)} = \boldsymbol{W}^t - \frac{1}{L_g}\nabla f\left(\boldsymbol{W}^{(t)}\right)$
4. $\boldsymbol{W}^{(t)} = \boldsymbol{W}_t + \frac{b_{t-1}-1}{b_t}(\boldsymbol{W}_t - \boldsymbol{W}_{t-1})$
5. $\boldsymbol{W}_{t+1} = \arg\min_{W} \varepsilon\|\boldsymbol{W}\|_1 + \frac{1}{2}\left\|\boldsymbol{W} - \boldsymbol{G}^{(t)}\right\|_F^2$
6. $b_{t+1} = \frac{1+\sqrt{4b_t^2+1}}{2}, t = t+1$
7. $t = t+1$
8. **end while**

B. Algorithm Pseudocode of Label Interactive Causal Inference Method

The label interactive causal inference method based on conditional independence test is summarized in Algorithm 2.

Algorithm 2: The Label Interactive Causal Inference Method based on Conditional Independence Test

Input: Label matrix: $\boldsymbol{S}$; Number of labels: l.

Output: Label interactive causality matrix: $\boldsymbol{V}$.

1. for $i = 1, 2, 3, ..., n$ do
2. for $j = 1, 2, 3, ..., n$ do
3. Construct $P(y_m)$ and $P(y_n \mid y_m)$, calculate $D_{y_n \to y_m} = D(P(y_n), P(y_m|y_n))$
   Construct $P(y_n)$ and $P(y_m \mid y_n)$, calculate $D_{y_m \to y_n} = D(P(y_m), P(y_m|y_n))$
4. end for
5. end for
6. If $\mathcal{D}_{y_n \to y_m} > \mathcal{D}_{y_m \to y_n}$, then $y_m \longrightarrow y_n$ is the inferred interactive causal direction, $\boldsymbol{V}_{mn} = 1$
   If $\mathcal{D}_{y_m \to y_n} > \mathcal{D}_{y_n \to y_m}$, then $y_n \longrightarrow y_m$ is the inferred interactive causal direction, $\boldsymbol{V}_{nm} = 1$
7. else, no causation, $\boldsymbol{V}_{mn} = 0$

C. Complexity Analysis

In summary, the algorithm complexity of CCSRMC is $(ndl(nd+n+d)+nl(l^2+n^3l+n)+dl^2)$. In order to comprehensively reflect the performance of CCSRMC, the algorithm complexity of this paper is compared with LSML [5], LSF-CI [3], FF-MLLA [6], LLSF [2] and ACML [8]. In comparison experiments, the closest comparison algorithm to the performance of this paper are LSML and FF-MLLA [6]. The algorithm complexity of LSML is $O((n+l)d^2+(n+d)l^2+ndl+d^3+l^3)$. FF-MLLA does not give a specific algorithm time complexity analysis. The algorithm complexity of LSF-CI is $O(nd^2+nd+ndl+lg^2+d^3+d^2l)$. Although the complexity of the algorithm in this paper is slightly higher than that of the comparison algorithms, the experimental results show that the CCSRMC algorithm has a better classification effect on most multi-label datasets than the comparison algorithms. Table I summarizes the computational complexity of the proposed methods and comparisons.

TABLE I. The Computational Complexity of Different Algorithm

Methods	Computational complexity
LSF-CI	$O(nd^2+nd+ndl+lg^2+d^3+d^2l)$
LSML	$O((n+l)d^2+(n+d)l^2+ndl+d^3+l^3)$
ACML	$O((n+l+nl)d^2+(n+d)l^2+3/2)$
LLSF	$O(d^2+dl+l^2+nd+nl)$
CCSRMC	$O(ndl(nd+n+d)+nl(l^2+n^3l+n)+dl^2)$

IV. Experiment

A. Dataset

To illustrate the effectiveness of the algorithm, 14 multi-label datasets from Yahoo.com and Mulan.com are selected. Table II is a detailed description. CCSRMC is a related model of multi-label classification, so in section IV-D, this paper selects multiple multi-label text classification datasets for comparison experiments. In order to reflect the universality of the algorithm in this paper, we also select other types of multi-label data sets for comparative experiments to compare and verify the effectiveness of the algorithm proposed in this paper.

TABLE II. Description of Datasets

Datasets	Train	Test	Labels	Features	Domain
Birds [2]	645	645	20	260	Image
Genbase[2]	662	662	27	1185	Biology
Enron[2]	1702	1702	53	1001	Text
Yeast[2]	2417	2417	14	103	Biology
Arts[1]	2000	3000	26	462	Text
Computers[1]	2000	3000	33	681	Text
Education[1]	2000	3000	33	550	Text
Science[1]	2000	3000	40	743	Text
Society[1]	2000	3000	27	636	Text
Entertainment[1]	2000	3000	21	640	News
Business[1]	2000	3000	30	438	News
Health[1]	2000	3000	32	612	Text
Reference[1]	2000	3000	33	793	Text
Recreation[1]	2000	3000	22	606	News

[1] Yahoo Web Pages (http://archive.ics.uci.edu/ml/)
[2] Mulan (http://mulan.sourceforge.net/datasets-mlc.html)

B. Comparison Algorithm and Parameter Settings

In this experiment, five multi-label classification algorithms are selected for comparison with CCSRMC. LSI-CI [3] is a multi-label classification algorithm that promotes label-specific features learning by learning correlation information between features and correlation information between labels. Its parameters are set to $\alpha=2^{10}$, $\beta=2^{8}$, $\gamma=1$, $\theta=2^{-8}$. LLSF [2] improves the performance of multi-label classification by learning the cosine similarity between labels to perform label-specific features learning. The parameters are set to $\alpha=2^{-4}$, $\beta=2^{-6}$, $\gamma=1$. LSML [5] handles the multi-label classification of the default data set by learning high-order label correlation matrix and label-specific features, and the parameters are set to $\lambda_2=10^{-5}$, $\lambda_3=10^{-3}$, $\lambda_4=10^{-5}$. FF-MLLA [6] uses the firefly method to fuse correlation information with sample similarity information. Then it classifies through singular value decomposition and extreme learning machine. In the FF-MLLA algorithm, the number of neighbors is k=15. The regularization coefficient is set to 1. The kernel function chooses RBF. The nuclear parameter is set to 100. The training method selects linear regression fitting. The ACML [8] algorithm uses cosine similarity to construct a label correlation matrix. Then the algorithm measures the adjacency between the labels to construct the label adjacency matrix. Finally, the label adjacency matrix is used to constrain the label correlation matrix to link the interactive causal label correlation. Its parameter setting interval is $\alpha\in[2^{-10},2^{10}]$, $\beta\in[2^{-10},2^{10}]$.

C. Metric

The evaluation index of the multi-label learning system is different from the traditional single-label learning system. The output label of the multi-label learning may be partially correct, completely correct, or completely wrong. In this paper, five evaluation indicators that are widely used in multi-label tasks are compared with the above-mentioned multi-label classification algorithms, including Hamming Loss, Average Precision, One-Error rate, Ranking Loss, AUC and Coverage rate [26] [27]. The value range of these evaluation indicators varies between [0,1]. For each evaluation indicator, "↑" means the larger, the better, and "↓" means the smaller, the better. Where $D=\{(X_{it},Y_{il}\mid 1\le t\le d, 1\le i\le m, 1\le l\le L)\}$ is multi-label dataset. $h(\cdot)$ is a multi-label classifier. $f(\cdot)$ is the prediction function. The definitions of 5 evaluation indicators are as follows:

Hamming Loss can be used to evaluate how many times a sample is misclassified. For example, a sample does not belong to label L_i but is incorrectly classified into label L_i. Or a sample belongs to label L_i but is not predicted as label L_i. The algorithm in this paper uses Hamming loss to calculate the numerical distance between the result sequence predicted by the classifier and the original result sequence.

$$HL_D(h)\downarrow=\frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i\Delta Z_i|}{M} \tag{36}$$

Where m is the number of samples, M is the total number of labels. Y_i is the set of actual labels of i-th sample, Z_i is the set of predicted labels of i-th sample. Δ refers to the symmetric difference between the two sets. The smaller the Hamming loss, the better the prediction result.

In the ranking of all prediction results, the average precision represents the probability that the ranking is ranked before the labels of the related label set and belongs to the related label set. The indicator reflects the average precision of the classification label. This indicator was originally used in Information Retrieval (*IR*) systems to evaluate the retrieval performance of text sorting.

$$AP_D(f)\uparrow=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|}\cdot\sum_{y\in Y_i}\frac{\left|\{y'\mid rank_f(x_i,y')\le rank_f(x_i,y),y'\in Y_i\}\right|}{rank_f(x_i,y)} \tag{37}$$

$rank_f$ is the ranking function. When the average precision reaches 1, the prediction effect is the best. That is, the larger the $AP_D(f)$, the higher the performance of $f(\cdot)$.

One-Error can be used to evaluate the probability that the label ranked first in the output result does not belong to the actual label set. It can reflect the times that the highest ranking object is incorrectly labeled.

$$OE_D(f) \downarrow = \frac{1}{m}\sum_{i=1}^{m} g\left[\left[arg\max_{y\in Y} f(x_i, y)\right] \notin Y_i\right],$$
$$g(x) = \begin{cases} 0 & x\ is\ false \\ 1 & x\ is\ \ true \end{cases} \tag{38}$$

The smaller the One-Error, the better the prediction. That is, the smaller the $OE_D(f)$, the higher the performance of $f(\cdot)$.

Ranking Loss indicates how many irrelevant labels are ranked higher than related labels. The ranking loss is used to indicate the average of the probability that a label that does not belong to the relevant label set is ranked in the relevant label set in the result ranking.

$$RL_D(f) \downarrow = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{|Y_i||\overline{Y}_i|} \cdot \left|\left\{\begin{matrix}(y_1, y_2)|f(x_i, y_1) \le f(x_i, y_2), \\ (y_1, y_2) \in Y_i \times \overline{Y}_i\end{matrix}\right\}\right| \tag{39}$$

The smaller the Ranking Loss, the better the prediction result. That is, the smaller the $RL_D(f)$, the higher the performance of $f(\cdot)$.

Coverage can be used to reflect the number of labels required to cover all labels in the label sequence of the evaluated object.

$$CV_D(f) \downarrow = \frac{1}{m}\sum_{i=1}^{m} \max_{y\in Y_i} rank_f(x_i, y) - 1 \tag{40}$$

The smaller the coverage, the better the prediction result, and the higher the performance of $f(\cdot)$.

AUC (Area under the Curve of ROC) is an evaluation index that measures the pros and cons of a two-class model. AUC represents the probability that a positive example is ranked before a negative example. When a positive sample and a negative sample are randomly selected, the current classification algorithm ranks the positive sample ahead of the negative sample according to the calculated score value. The larger the AUC, the more likely the current classification algorithm will rank the positive samples in front of the negative samples. Therefore, the effect of classification is better.

$$AUC_D(f) \uparrow = \frac{\sum_{i\in positiveClass}^{m} rank_i - \frac{M(1+M)}{2}}{M \times N} \tag{41}$$

The area under the ROC curve is between 0.1 and 1. As a value, AUC can intuitively evaluate the quality of the classifier, the larger the value, the better the effect.

V. Analysis and Visualization

A. Analysis of the Results

In the experiment, this article uses a five-fold cross-validation method to evaluate the performance of the algorithm. Five-fold cross-validation means that all data is randomly divided into five equal subsets, each subset is tested in turn, and the remaining data is used for training. Five-fold cross-validation is iterated five times, so the average value after five runs of the experiment needs to be calculated. The five comparison algorithms selected in this article all consider different symmetric label correlations.

As illustrated in Table Ⅲ, the experimental results of each algorithm under 14 datasets and the optimal experimental results have been marked in bold-type. Analyzing the above experimental results, we get the following conclusions:

1. It can be seen from Table Ⅲ that in the 84 sets of experimental data, CCSRMC has the best results in 53 sets, with a dominant ratio of 63%. The evaluation index AP is significantly better than other comparison algorithms. The performance of 13 out of 14 data sets is better than other comparison algorithms. In addition, the variance of the CCSRMC algorithm is smaller, which shows that the performance of CCSRMC is more stable.
2. Compared with the ACML algorithm that also takes the interactive causal correlation between labels into account, the overall performance of CCSRMC is better than the ACML algorithm. The reason is that although the two algorithms both consider the interactive causal correlation between labels, the CCSRMC algorithm uses the label space reshaping method to transform the original discrete labels into continuous labels. The use of continuous labels to infer the interactive causal correlation between features and labels cause the results of the CCSRMC algorithm be superior to the ACML algorithm to a certain extent.
3. The algorithm LSML combines high-order label correlation matrix and specific features to process the multi-label classification of the default data set. In the five indexes of HL, OE, RL, AUC and CV, the experimental results show that the algorithm CCSRMC proposed in this paper is significantly better than the algorithm LSML, which verifies the effectiveness of the algorithm in this paper. Thus, by considering the interactive causal relationship between labels, different labels with dependencies can be better identified and the redundant information in the model is reduced, which can improve the performance of the multi-label classifier to a certain extent.

B. Ablation Analysis

In order to verify that the introduction of interactive causal label correlation in the model improves the performance of the algorithm, we conduct related experiments for ablation analysis in this section. We compared CCSRMC using an interactive causal label correlation matrix with SRMC using a label correlation matrix. Some results are shown in Fig. 3. The performance of the CCSRMC algorithm using the interactive causal label correlation matrix is better than that of the CCSRMC algorithm using the label correlation matrix. It further illustrates the effectiveness and rationality of introducing interactive causal label correlation in the multi-label algorithm.

C. Parameter Sensitivity Analysis

The algorithm CCSRMC proposed in this paper has three parameters α, β, γ. The parameter α controls the influence of the interactive causal correlation between the labels on the model coefficients and the weight constraints that minimize structural risks and prevent overfitting. The parameter β controls the sparseness of label features extracted from the label-specific features. The parameter γ controls the sparsity of the numerical label matrix S. By a method that fix two parameter values and change one parameter value to find the optimal value, we found that the parameter β has no obvious change in the six evaluation indicators. This further verifies that we did not use $\boldsymbol{l}_1$-norm regularization to extract specific features, but use interactive causal correlation to connect specific features between labels. In this section, we use Bar-3 to visualize the parameter sensitivity comparison of parameters α and γ. The algorithm in this paper conducts parameter sensitivity experiments on the Emotions data set. According to the experimental results in Fig. 4, it can be found that the algorithm CCSRMC has different sensitivity to the regularization parameters on the six evaluation indicators. On the evaluation index of AUC, when the experimental interval of parameter α is set to $[2^7, 2^{10}]$, the parameter γ is affected, which leads to the deterioration of the performance of the algorithm. For the HL evaluation index, as the parameter α interval increases, the HL index decreases and then rises. When the parameter $\alpha > 2^6$, the correlation information between the labels obtained by the CCSRMC algorithm becomes very scarce, and the risk of the model structure increases, which may easily lead to overfitting. When the parameter $\gamma > 2^2$, the label specific features that can be extracted in the CCSRMC algorithm become very scarce, and

TABLE III-A. TEST RESULTS OF EACH ALGORITHM ON 6 EVALUATION INDEXES (MEAN ± STD)

Dataset	Metric	CCSRMC	ACML	FF-MLLA	LSML	LLSF	LSF-CI
Birds	HL↓	**0.0475**±0.0029	0.0512±0.0023	0.0476±0.0040	0.0595±0.0046	0.0506±0.0041	0.0647±0.0063
	AP↑	**0.7687**±0.0304	0.7648±0.0275	0.7517±0.0247	0.7596±0.0219	0.7582±0.0350	0.6302±0.0165
	OE↓	**0.2759**±0.0358	0.2821±0.0301	0.3007±0.0358	0.2884±0.0458	0.2915±0.0385	0.4062±0.0233
	RL↓	**0.0887**±0.0143	0.0891±0.0150	0.1067±0.0116	0.0972±0.0061	0.0962±0.0232	0.2206±0.0168
	CV↓	**0.1382**±0.0192	0.1391±0.0210	0.1512±0.0206	0.1483±0.0162	0.1471±0.0322	0.2739±0.0156
	AUC↑	**0.9112**±0.0143	0.8953±0.046	0.7714±0.0145	0.6498±0.0040	0.7690±0.0178	0.6978±0.0145
Arts	HL↓	**0.0522**±0.0007	0.0536±0.0007	0.0588±0.0015	0.0582±0.0011	0.0566±0.0009	0.0561±0.0013
	AP↑	**0.6305**±0.0072	0.6241±0.0141	0.5211±0.0101	0.5932±0.0069	0.5852±0.0147	0.5451±0.0100
	OE↓	**0.4436**±0.0087	0.4524±0.0179	0.607±0.0191	0.4762±0.0088	0.4900±0.0181	0.5090±0.0180
	RL↓	**0.1350**±0.0070	0.1405±0.0074	0.1571±0.0031	0.1770±0.0058	0.1841±0.0106	0.2621±0.0106
	CV↓	**0.2065**±0.0102	0.2141±0.0080	0.2212±0.0050	0.2567±0.0060	0.2650±0.0117	0.3448±0.0107
	AUC↑	**0.8649**±0.0070	0.7832±0.0891	0.5558±0.0059	0.6723±0.0178	0.6916±0.0114	0.7102±0.0007
Computers	HL↓	**0.0326**±0.0006	0.0339±0.0009	0.0383±0.0010	0.0391±0.0008	0.0389±0.0010	0.0415±0.0018
	AP↑	**0.7174**±0.0034	0.7093±0.0163	0.6424±0.0041	0.6915±0.0059	0.6575±0.0064	0.5839±0.0059
	OE↓	**0.3372**±0.0049	0.3466±0.0171	0.4302±0.0072	0.3608±0.0060	0.4080±0.0094	0.4614±0.0072
	RL↓	0.0903±0.0043	**0.0980**±0.0086	0.0974±0.0044	0.1230±0.0059	0.1229±0.0059	0.2299±0.0126
	CV↓	0.1330±0.0064	**0.1406**±0.0106	0.1402±0.0045	0.1725±0.0051	0.1720±0.0088	0.2888±0.0170
	AUC↑	**0.9096**±0.0043	0.7828±0.0916	0.6764±0.0041	0.7505±0.0165	0.6810±0.0136	0.7813±0.0068
Education	HL↓	0.0381±0.0003	**0.0371**±0.0008	0.0407±0.0002	0.0411±0.0003	0.0414±0.0007	0.0418±0.0012
	AP↑	**0.6670**±0.0067	0.6337±0.0153	0.5497±0.0050	0.6033±0.0082	0.5805±0.0069	0.5290±0.0166
	OE↓	0.4666±0.0083	**0.4606**±0.0203	0.5868±0.0079	0.4826±0.0178	0.5090±0.0070	0.5290±0.0166
	RL↓	**0.0990**±0.0056	0.1089±0.0057	0.1001±0.0053	0.1526±0.0068	0.1642±0.0065	0.2486±0.0081
	CV↓	**0.1444**±0.0053	0.1592±0.0086	0.1323±0.0067	0.2123±0.0077	0.2215±0.0068	0.3133±0.0115
	AUC↑	**0.9009**±0.0056	0.8709±0.0190	0.5612±0.0037	0.6435±0.0003	0.6660±0.0160	0.6784±0.0532
Enron	HL↓	0.0464±0.0005	**0.0455**±0.0006	0.0500±0.0007	0.0505±0.0014	0.0553±0.0020	0.0664±0.0066
	AP↑	**0.7061**±0.0083	0.7027±0.0099	0.6546±0.0052	0.6928±0.0088	0.6609±0.0085	0.6539±0.0060
	OE↓	0.2297±0.0179	**0.2279**±0.0216	0.2584±0.0139	0.2444±0.0156	0.2650±0.0147	0.2644±0.0114
	RL↓	**0.0841**±0.0061	0.0895±0.0039	0.1000±0.0032	0.0951±0.0043	0.1258±0.0059	0.1109±0.0023
	CV↓	**0.2430**±0.0122	0.2547±0.0072	0.2753±0.0083	0.2695±0.0100	0.3154±0.0156	0.2779±0.0045
	AUC↑	**0.9158**±0.0061	0.8367±0.0037	0.7053±0.0051	0.6329±0.0089	0.6673±0.0094	0.6307±0.0275
Genbase	HL↓	0.0023±0.0007	**0.0015**±0.0004	0.9540±0.0014	0.0010±0.0005	0.4130±0.2064	0.0040±0.0012
	AP↑	0.9946±0.0029	0.9952±0.0034	0.9926±0.0061	**0.9969**±0.0035	0.3133±0.1597	0.9966±0.0034
	OE↓	0.0015±0.0030	0.0015±0.0030	——	0.0015±0.0031	0.0061±0.2239	0.0015±0.0030
	RL↓	0.0022±0.0020	0.0029±0.0025	0.0050±0.0048	**0.0016**±0.0028	0.0645±0.0300	0.0018±0.0030
	CV↓	0.0131±0.0049	0.0123±0.0011	0.0177±0.0084	0.0125±0.0045	0.0809±0.2143	**0.0101**±0.0016
	AUC↑	**0.9977**±0.0020	0.8900±0.0039	0.6849±0.0288	0.7106±0.0044	0.7683±0.0303	0.5903±0.0013
Entertainment	HL↓	0.0532±0.0010	**0.0508**±0.0014	0.0589±0.0005	0.0570±0.0006	0.0550±0.0015	0.0550±0.0014
	AP↑	**0.6944**±0.0107	0.6925±0.0067	0.5777±0.0110	0.6731±0.0089	0.6669±0.0071	0.6351±0.0076
	OE↓	0.3948±0.0109	**0.3912**±0.0106	0.5668±0.0178	0.4072±0.0113	0.4092±0.0070	0.4166±0.0063
	RL↓	**0.1089**±0.0068	0.1163±0.0035	0.1284±0.0047	0.1422±0.0040	0.1460±0.0110	0.2215±0.0096
	CV↓	**0.1523**±0.0085	0.1605±0.0066	0.1661±0.0052	0.1897±0.0044	0.1918±0.0114	0.2717±0.0109
	AUC↑	**0.8910**±0.0068	0.8013±0.0009	0.5879±0.0035	0.6128±0.0034	0.7600±0.1127	0.5699±0.0901

TABLE III-B. TEST RESULTS OF EACH ALGORITHM ON 6 EVALUATION INDEXES (MEAN ± STD)

Dataset	Metric	CCSRMC	ACML	FF-MLLA	LSML	LLSF	LSF-CI
Business	HL↓	0.0278±0.0007	**0.0266**±0.0004	0.0261±0.0016	0.0287±0.0010	0.0295±0.0009	0.0398±0.0008
	AP↑	**0.8838**±0.0055	0.8809±0.0072	0.8805±0.0110	0.8798±0.0085	0.8484±0.0090	0.7825±0.0116
	OE↓	0.1176±0.0106	**0.1158**±0.0074	0.114±0.0101	0.1104±0.0105	0.1452±0.0112	0.2240±0.0185
	RL↓	**0.0407**±0.0018	0.0443±0.0048	0.0452±0.0049	0.0485±0.0048	0.0635±0.0041	0.1036±0.0071
	CV↓	**0.0818**±0.0045	0.0895±0.0078	0.0833±0.0068	0.0967±0.0097	0.1096±0.0064	0.1559±0.0092
	AUC↑	**0.9592**±0.0018	0.8971±0.0142	0.8520±0.0075	0.7990±0.0005	0.7168±0.0093	0.7009±0.0141
Science	HL↓	0.0322±0.0009	**0.0311**±0.0006	0.0348±0.0008	0.0333±0.0007	0.0348±0.0007	0.0348±0.0007
	AP↑	**0.6041**±0.0097	0.6077±0.0081	0.4556±0.0135	0.5890±0.0158	0.5521±0.0105	0.5166±0.0116
	OE↓	0.4846±0.0148	**0.4772**±0.0108	0.6694±0.0146	0.4884±0.0207	0.5274±0.0075	0.5512±0.0191
	RL↓	**0.1171**±0.0061	0.1289±0.0046	0.1568±0.0025	0.1530±0.0106	0.1774±0.0083	0.2473±0.0056
	CV↓	**0.1625**±0.0129	0.1778±0.0058	0.1978±0.0057	0.2048±0.0143	0.2296±0.0087	0.3027±0.0065
	AUC↑	0.8828±0.0061	0.8600±0.0003	0.5346±0.0028	0.6415±0.0340	**0.8962**±0.0031	0.7624±0.0007
Society	HL↓	0.0526±0.0020	**0.0515**±0.0010	0.0561±0.0015	0.0560±0.0010	0.0578±0.0011	0.0585±0.0013
	AP↑	**0.6397**±0.0141	0.6360±0.0076	0.5872±0.0126	0.6128±0.0118	0.5930±0.0071	0.5158±0.0108
	OE↓	0.3946±0.0183	**0.3900**±0.0106	0.4634±0.0178	0.4028±0.0183	0.4462±0.0099	0.5192±0.0192
	RL↓	**0.1407**±0.0069	0.1517±0.0061	0.1503±0.0051	0.1879±0.0054	0.1857±0.0042	0.2820±0.0071
	CV↓	**0.2245**±0.0116	0.2387±0.0115	0.2274±0.0069	0.2847±0.0064	0.2755±0.0061	0.3820±0.0090
	AUC↑	**0.8592**±0.0069	0.7723±0.0087	0.6280±0.0066	0.7098±0.0037	0.6449±0.0083	0.7222±0.0056
Yeast	HL↓	0.2055±0.0038	**0.2000**±0.0033	0.1955±0.0072	0.2608±0.0082	0.2008±0.0036	0.0585±0.0013
	AP↑	0.7504±0.0109	**0.7570**±0.0067	0.7626±0.0089	0.6102±0.0082	0.7587±0.0091	0.5158±0.0108
	OE↓	0.2407±0.0175	**0.2291**±0.0146	0.2304±0.0106	0.3583±0.0192	0.2275±0.0134	0.5192±0.0192
	RL↓	0.1727±0.0070	**0.1706**±0.0019	0.1717±0.0078	0.3463±0.0122	0.1696±0.0054	0.2820±0.0071
	CV↓	**0.4510**±0.0057	0.4534±0.0041	0.4543±0.0115	0.6233±0.0131	0.4530±0.0053	0.3820±0.0090
	AUC↑	**0.8272**±0.0070	0.7823±0.0010	0.7499±0.0089	0.6381±0.0107	0.6710±0.0036	0.6563±0.0096
Health	HL↓	**0.0301**±0.0012	0.0331±0.0008	0.0407±0.0011	0.0413±0.0006	0.0356±0.0010	0.0389±0.0009
	AP↑	**0.7840**±0.0073	0.7821±0.0062	0.7069±0.0088	0.7646±0.0098	0.7571±0.0059	0.6980±0.0094
	OE↓	0.2578±0.0138	**0.2510**±0.0068	0.3776±0.0121	0.2644±0.0109	0.2764±0.0141	0.3172±0.0162
	RL↓	**0.0609**±0.0040	0.0677±0.0041	0.0685±0.0045	0.0876±0.0044	0.0905±0.0034	0.1651±0.0060
	CV↓	**0.1183**±0.0078	0.1286±0.0072	0.1168±0.0074	0.1591±0.0060	0.1562±0.0035	0.2456±0.0067
	AUC↑	**0.9390**±0.0040	0.7906±0.0019	0.7011±0.0046	0.6954±0.0200	0.7199±0.0011	0.6435±0.0042
Reference	HL↓	**0.0004**±0.0268	0.0257±0.0008	0.0292±0.0004	0.0294±0.0010	0.0280±0.0006	0.0298±0.0014
	AP↑	0.0095±0.7090	**0.7135**±0.0033	0.6301±0.0078	0.7052±0.0072	0.6634±0.0129	0.5929±0.0173
	OE↓	**0.0122**±0.3760	0.3642±0.0090	0.4658±0.0084	0.3666±0.0088	0.4020±0.0144	0.4692±0.0178
	RL↓	**0.0045**±0.0891	0.0930±0.0047	0.0934±0.0047	0.1070±0.0060	0.1398±0.0087	0.2426±0.0149
	CV↓	**0.0069**±0.1134	0.1194±0.0079	0.1100±0.0057	0.1354±0.0072	0.1705±0.0104	0.2745±0.0171
	AUC↑	——	**0.7505**±0.056	0.6461±0.0035	0.7234±0.0093	0.6983±0.0023	0.6728±0.0003
Recreation	HL↓	0.0559±0.0014	**0.0535**±0.0011	0.9361±0.0016	0.0578±0.0008	0.0571±0.0013	0.0565±0.0003
	AP↑	**0.6436**±0.0125	0.6391±0.0043	0.4892±0.0039	0.6185±0.0097	0.5985±0.0148	0.5692±0.0097
	OE↓	0.4492±0.0199	**0.4444**±0.0092	0.6616±0.0071	0.4614±0.0102	0.4890±0.0248	0.5056±0.0113
	RL↓	**0.1416**±0.0027	0.1485±0.0037	0.1830±0.0015	0.1741±0.0107	0.1868±0.0045	0.2446±0.0075
	CV↓	**0.1898**±0.0048	0.1992±0.0059	0.2221±0.0028	0.2277±0.0123	0.2392±0.0044	0.2968±0.0088
	AUC↑	**0.8583**±0.0055	0.8022±0.0071	0.5339±0.0017	0.6991±0.0189	0.7764±0.0080	0.6101±0.0019

Note: "↑" ("↓") means that the larger (smaller) the evaluation index is, the better. The best results are displayed in bold magnification.

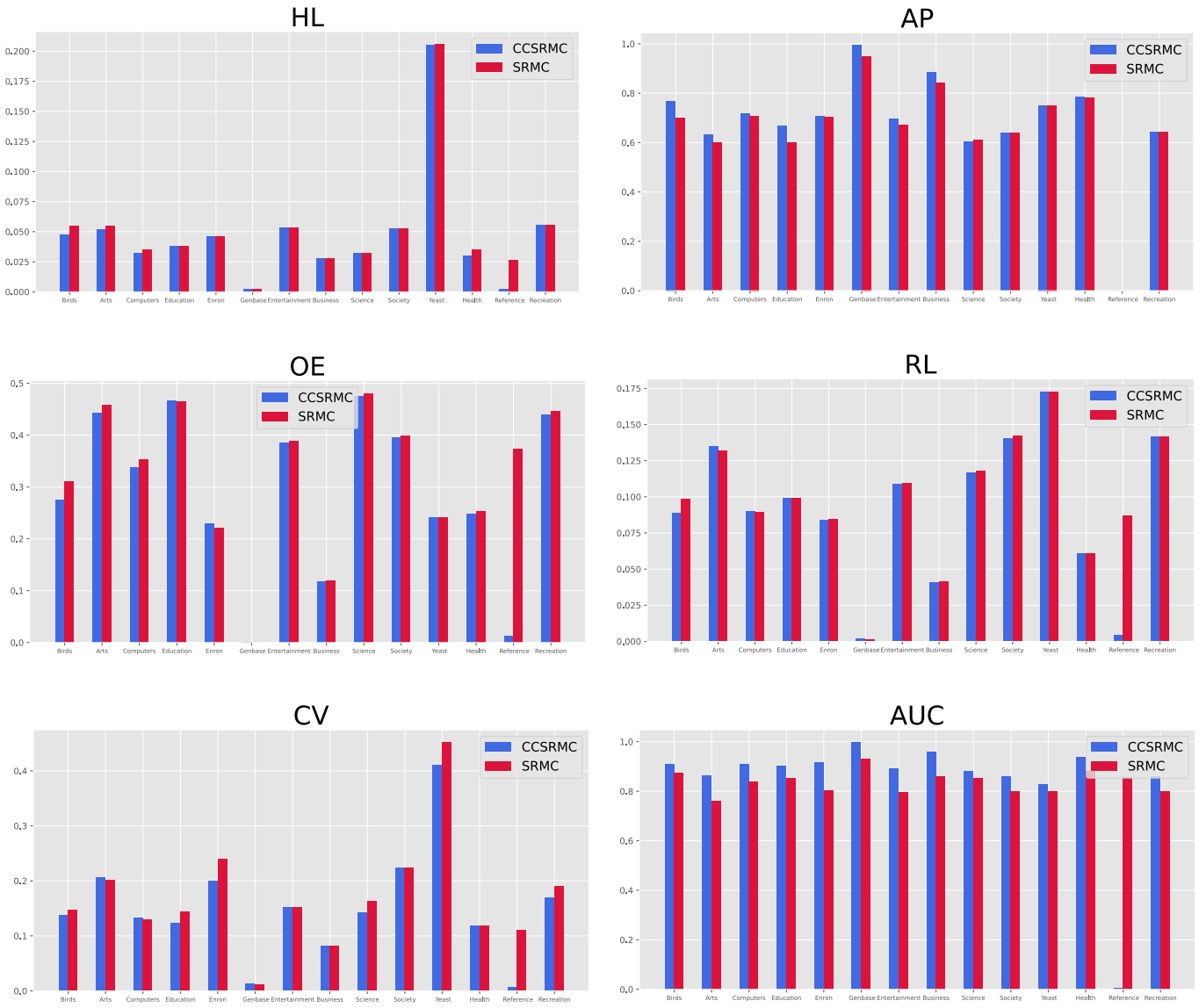

Fig. 3. Ablation Analysis of CCSRMC and SRMC on 6 Evaluation Indexes.

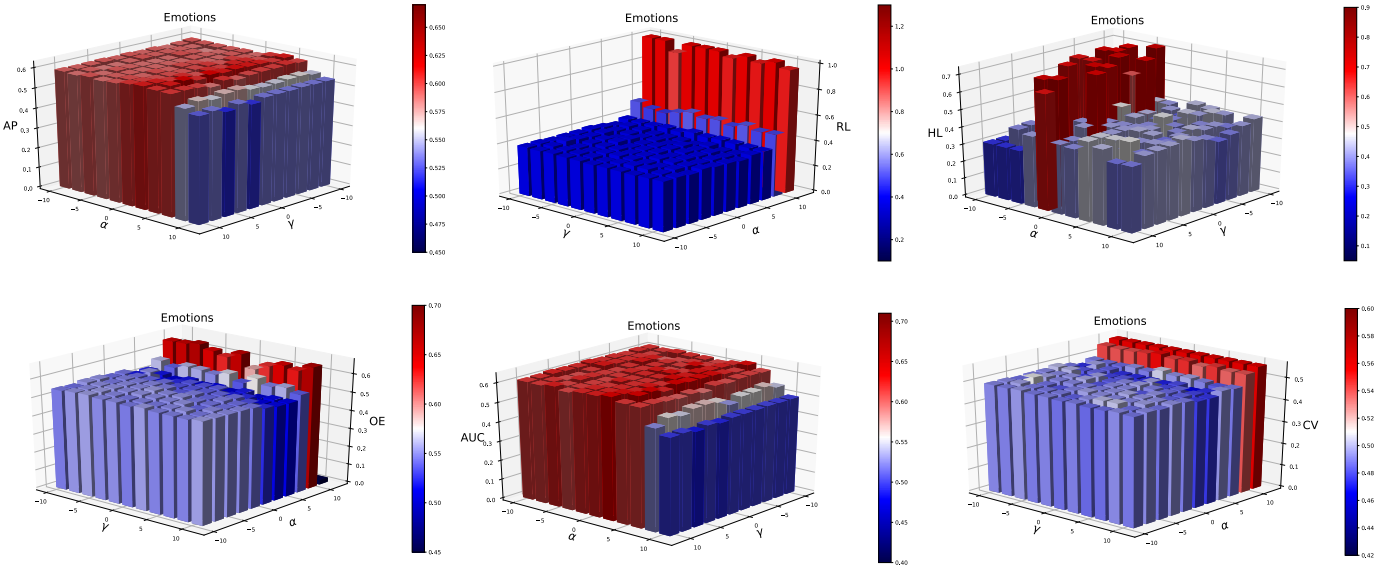

Fig. 4. Parameter Sensitivity Analysis Units.

the learning effect of label-specific features decreases. For AP and CV evaluation indexes, it can be found that the changes of the parameters α and γ have a relatively stable influence on the algorithm. In the RL evaluation index, we observe that when the parameter $\alpha > 2^8$, the performance of the algorithm drops rapidly. In the OE evaluation index, we found that as the parameter α increases, the performance of the algorithm first increases and then decreases. In the interval $[2^{-10}, 2^0]$, the performance of CCSRMC algorithm is relatively stable. Combining the sensitivity analysis of each parameter above, it is suggested that the parameter setting interval in this paper is $\alpha\epsilon[2^{-10}, 2^{-1}]$, $\beta\epsilon[2^{-10}, 2^{10}]$, $\gamma\epsilon[2^{-10}, 2^6]$.

D. Statistical Hypothesis Testing References and Footnotes

In this paper, the stability of the performance of CCSRMC and other comparative experimental algorithms on 14 datasets is compared by using the Nemenyi test [28] with a significance level of 5%. When the average ranking difference of the two comparison algorithms on all data sets is greater than the critical difference (*CD*), it is considered that there is a significant difference between the two algorithms, otherwise it is considered that there is no significant difference. The calculation formula of *CD* value is:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{42}$$

Where the significance level is $\alpha = 0.05$, $k = 6$, $N = 14$, $q_\alpha = 2.850$, so $CD = 2.2518$.

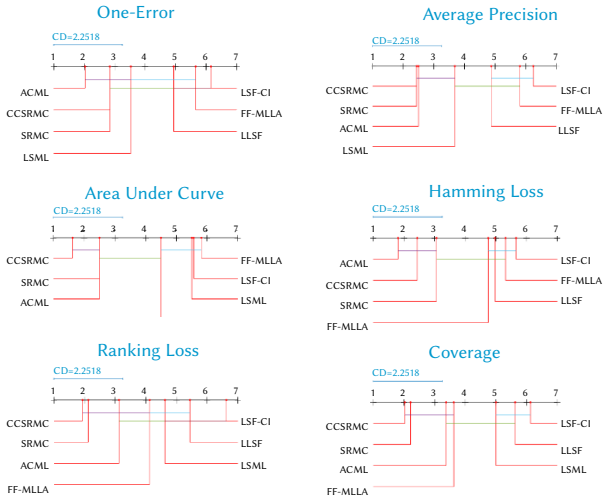

Fig.5. Parameter Sensitivity Analysis Units.

According to the results shown in Fig. 5. The higher the average ranking, the more dominant the algorithm. Compared with other comparison algorithms, the average rankings of CCSRMC are all the best on the four evaluation indexes of AP, AUC, RL, and CV. It is slightly inferior to the ACML algorithm on the OE and HL indexes. The average ranking of SRMC, which does not consider the interactive causal relationship between labels, is always inferior to CCSRMC. The validity and rationality of introducing interactive causal label correlation in the multi-label algorithm is again verified. In the OE index, the CCSRMC algorithm is significantly different from LSF-CI, FF-MLLA, and ACML. The average accuracy of CCSRMC is significantly different from LLSF, LSF-CI, and FF-MLLA. In the indexes of AUC and HL, CCSRMC is significantly different from other comparison algorithms except ACML. In the indexes of AUC and HL, CCSRMC is significantly different from other comparison algorithms except ACML. In terms of RL and CV evaluation indexes, LLSF, LSF-CI, LSML are all significantly different from CCSRMC. The results of Nemenyi test are consistent with the basic results of experimental analysis. The results of the Nemenyi test further verify the performance of the algorithm in this paper, which shows that the introduction of interactive causal inference in multi-label learning is reasonable and effective.

VI. Conclusion

This article is using the spatial reshaping method to transform the original discrete label into a continuous label. On the basic of solving the problems that the existence of logical label cannot describe the importance of different labels and cannot fully represent semantic information, the label co-occurrence matrix is constructed by combining the conditional probability test method. The label co-occurrence matrix and the label correlation matrix are combined to construct the label interactive causal correlation matrix to perform multi-label classification learning on the obtained numerical label matrix. It avoids the problem that the distance failed to measure high-dimensional space and the parameter depends on manual selection. The experimental results show that the method has a certain validity. The accuracy of multi-label classification is improved. What's more, the interactive causal situation of the correlation between the labels is considered to reduce the redundant information in the classification model. However, the method we proposed still needs improvement. For example, the problem of missing labeling and wrong labeling caused by the default of label data may affect the accuracy of interactive causal inference. The method in this paper only considers the dependency relationship between paired variables (a set of labels or labels and features), while ignoring the influence of factors such as confounding variables. The experimental results on the image data set show that only considering the dependency between paired variables is not suitable for more complex scenarios. The use of continuous labels for training should cooperate with an appropriate dynamic threshold selection mechanism. Each comparison algorithm does not fully consider the label distribution of each sample. Although the introduction of interactive causal inference in multi-label learning has achieved certain results, the method adopted is relatively simple. Thus, further study and research are necessary.

Acknowledgment

This work was supported by the National Natural Science Foundation of Anhui under Grant 2108085MF216 and Key Laboratory of Data Science and Intelligence Application, Fujian Province University (NO. D202005).Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education (Anhui University) (No.2020A003) and Anqing Normal University Graduate Innovation Fund(No.2021yjsXSCX017).

References

[1] Y. B. Wang, W. J. Zheng, Y. S. Cheng, T. C. Cao, "Multi-label classification algorithm based on PLSA learning probability distribution semantic information", *Journal of Nanjing University (Natural Science)*, vol.57, no.1, pp.75-89, 2021.

[2] J. Huang, F. Qin, X. Zheng, "Learning label-specific features for multi-label classification", *Information Sciences: An International Journal*, vol. 492, no.18, pp.124-146, 2019.

[3] H. R. Han, M. X. Huang, Y. Zhang, X. G. Yang, W. G. Feng, "Multi-label learning with label specific features using correlation information", *IEEE Access*, vol.19, no.7, pp. 11474–11484, 2019.

[4] M. L. Zhang, L. Wu, "Multi-label learning with label-specific features", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.37,

no.1, pp. 107–120, 2015.
[5] J. Huang, F. Qin, X. Zheng, "Improving multi-label classification with missing labels by learning label-specific features", *Information Sciences: An International Journal*, vol.492, pp.124–146, 2019.
[6] Y. S. Cheng, K. Qian, Y. B. Wang, D. W. Zhao, "Multi-label lazy learning approach based on firefly method", *Journal of Computer Applications*, vol.39, no.5, pp.1305–1311, 2019.
[7] G. M. Zhang, B. Y. Long, J. X. Zeng, J. Y. Huang, "Zero-shot attribute recognition based on de-redundancy features and semantic relationship constraint", *Pattern Recognition and Artificial Intelligence*, vol.34, no.9, pp.809-823, 2021.
[8] J. C. Bao, Y. B. Wang, Y. S. Cheng, "Asymmetry label correlation for multi-label learning", *Applied Intelligence*. Prepublish (2021), doi:10. 1007/S10489-021-02725-4.
[9] D. D. Lai, Z. H. Luo, Y. L. Ma, "Label order optimization method of classifier chains based on co-occurrence analysis", *Systems Engineering and Electronics*, vol.43, no.9, pp.2526-2534, 2021.
[10] C. X. Yan, S. G. Zhou, "Effective and scalable causal partitioning based on low-order conditional independent tests", *Neurocomputing*, vol.389, no.14, pp.146-154, 2020.
[11] X. S. Yin, T. Shu, Q. Huang, "Semi-supervised fuzzy clustering with metric learning and entropy regularization", *Knowledge-based systems*, vol.35, pp.304-311, 2012.
[12] W. B. Qian, Y. S. Xiong, J. Yang, W. H. Shu, "Feature selection for label distribution learning via feature similarity and label correlation" *Information Sciences*, vol.582, pp.38–59, 2022.
[13] Y. S. Cheng, C. Zhang, S. F. Pang, "Multi-label space reshape for semantic-rich label-specific features learning", *International Journal of Machine Learning and Cybernetics*, vol.13, pp.1005–1019, 2022, doi:10.1007/s13042-021-01432-3.
[14] A. Kale, Y. F. Wu, J. Hullman, "Causal support: modeling causal inferences with visualizations", *IEEE transactions on visualization and computer graphics*, 2021, doi:10.1109/TVCG.2021. 3114824.
[15] X. J. Song, A. Taamouti, "Measuring granger causal ity in quantiles", *Journal of Business & Economic Statistics*, vol.39, no.4, pp.1-48, 2021.
[16] De La Pava Panche Iván, Álvarez Meza Andrés, Herrera Gómez Paula Marcela, Cárdenas Peña David, Ríos Patiño Jorge Iván, Orozco Gutiérrez Álvaro, "Kernel-based phase transfer entropy with enhanced feature relevance analysis for brain computer interfaces", *Applied Sciences*, vol.11, no.15, pp.6689-6689, 2021.
[17] Z. C. Sha, Z. M. Liu, C. Ma, J. Chen, "Feature selection for multi-label classification by maximizing full-dimensional conditional mutual information", *Applied Intelligence*, vol.51, no.1, pp.326-340, 2020.
[18] X. Y. Li, H. L. Wang, B. Y. Wu, "A stable and efficient technique for linear boundary value problems by applying kernel functions", *Applied Numerical Mathematics*, vol.172, no.1, pp.206-214, 2022.
[19] G. J. Székely, M. L. Rizzo, N. K. Bakirov, "Measuring and testing dependence by correlation of distances", *The Annals of Statistics*, vol.35, pp.2769–2794, 2007
[20] R. C. Cai, Y. M. Bai, J. Qiao, Z. F. Hao, "Causal inference method based on confounder hidden compact representation model", J*ournal of Computer Applications*, 2021, doi:10.11772/j.issn.1001-9081.2020 122066.
[21] Z. H. Zhou, M. L. Zhang, "Multi-label learning", *Encyclopedia of Machine Learning and Data Mining*, *Berlin*, 2016, pp.875–881, Springer.
[22] Y. B. Wang, W. J. Zheng, Y. S. Cheng, D. W. Zhao, "Joint label completion and label-specific features for multi-label learning algorithm", *Soft Computing*, vol.24, pp.6553–6569, 2020.
[23] A. Beck, M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems", *IEEE transactions on image processing*, vol.18, no.11, pp.2419–2434, 2009.
[24] G. C. Liu, Z. C. Lin, Y. Yu, "Robust subspace segmentation by low-rank representation", *Proceeding Twenty-Seventh International Conference on Machine Learning*, 2010, pp.663–670.
[25] Z. C. Lin, A. Ganesh, J. Wright, L. Q. Wu, M. Chen, Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix", *UIUC Technical Report* 2009, vol.09, pp.2214.
[26] Z. H. Zhou, M. L. Zhang, "A Review on multi-label learning algorithms", *IEEE Transactions on Knowledge and Data Engineering*, vol.26, no.8, pp.1819–1837, 2014.
[27] W. Weng, Y. J. Lin, Y. W. Li, "Online multi-label streaming feature selection based on neighborhood rough set", *Pattern Recognition: The Journal of the Pattern Recognition Society*, vol.84, no.1, pp.273–287, 2018.
[28] J. Demiar, D. Schuurmans, "Statistical comparisons of classifiers over multiple datasets", *Journal of Machine Learning Research*, vol.7, no.1, pp.1–30, 2006.

Chao Zhang

Graduate student of computer and Information College, Anqing Normal University. His main research includes machine learning, data mining and statistics.

Yusheng Cheng

Professor of computer and Information College, Anqing Normal University. He received his Ph.D. in the School of Computer and Information Science of Hefei University of Technology in 2007. His research interests concern the rough set theory and algorithm, semi supervised learning and data mining. He is the author of more than 50 papers in journals and conference proceedings such as Information Science, Knowledge-Based Systems, Neurocomputing, Applied Soft Computing, PAKDD and so on.

Yibin Wang

Professor of computer and Information College, Anqing Normal University. The main research directions include multi label learning, machine learning and software security.

Yuting Xu

Graduate student of computer and Information College, Anqing Normal University. Her main research includes machine learning, data mining and statistics.