Parallel Dual-channel Multi-label Feature Selection

Jiali Miao[1], Yibin Wang[1,2], Yusheng Cheng[1,2,*], Fei Chen[3]

[1] School of Computer and Information, Anqing Normal University, Anhui, Anqing, 246011, China; 1150838286@qq.com (J.M.); wybjwqwy@163.com (Y.W.)
[2] The University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anhui, Anqing, 246133, China;
[3] School of Mathematics and Computer, Tongling University, Anhui, Tongling, 244061, China; 897204638@qq.com (F.C.)
* Correspondence: chengyshaq@163.com

Abstract: In the process of multi-label learning, feature selection methods are often adopted to solve the high-dimensionality problem in feature spaces. Most existing multi-label feature selection algorithms focus on exploring the correlation between features and labels and then obtain the target feature subset by importance ranking. These algorithms commonly use single-channel structure to obtain important features, which induces the excessive reliance on the ranking results and causes the loss of important features. However, the correlation between label-specific feature and label-instance is ignored. Therefore, this paper proposes Parallel Dual-channel Multi-label Feature Selection algorithm (PDMFS). We first introduce the concept of dual channel and design the algorithm model as two independent modules. The algorithm obtained different feature correlation sequences, thus avoided relevant feature loss. And then, the proposed algorithm uses the subspace model to select the feature subset with the maximum correlation and minimum redundancy for each sequence, thus obtaining feature subsets under respective correlations. Finally, the subsets are cross-merged to reduce the important feature loss caused by the serial structure processing single feature correlation. The experimental results on eight datasets and statistical hypothesis testing indicate that the proposed algorithm is effective.

Keywords: Multi-label learning; Feature selection; Parallel dual-channel; Information entropy; Cross-merging

1. Introduction

As a research hotspot in the fields of machine learning, pattern recognition, and data mining, multi-label learning [1] has attracted much attention. In contrast to traditional single-label learning, multi-label learning can handle classification tasks with multiple targets, which is more relevant to realistic classification tasks. However, with the advent of 5G information era, multi-label data gradually becomes a kind of huge and high-dimensional data. However, the high-dimensional features of multi-label datasets inevitably include irrelevant or redundant features, which can lead to dimensional disasters and reduce the efficiency of multi-label classification tasks. Feature selection is an effective technique to solve the high-dimensional disaster of data [2,3]. It reduces feature dimensionality by selecting feature subsets with the least number of irrelevant or redundant features with a certain feature-specific metric [4-6].

In recent years, various multi-label feature selection algorithms [7-10] have been proposed. These algorithms are usually divided into three main categories: filtering methods, wrapper methods and embedding methods. The filtering methods perform classifier-independent feature selection, and it measures the data features, such as mutual information, and produces a feature ranking before classification; the wrapper methods use the accuracy of a particular classifier to determine the selected feature quality, it rely on the learning performance of an off-the-shelf classifier for the evaluation of selected features; and the embedding methods use feature selection as a function component of a predefined classifier and achieve model fitting and feature selection simultaneously. In this paper, we consider only filtering methods to design efficient evaluation metrics and evaluation methods for feature selection.

Current multi-label feature selection algorithms differ in their approach to selecting feature subsets, but the goals are extremely similar, i.e., first performing relevance analysis on features and then selecting features based on established requirements (the requirement is usually to select low-redundancy features). If the relevance analysis and selection requirements are modularized, the general structure of current algorithms is serial connecting modules, that is called a single-channel structure in this paper. Single-channel structure has achieved remarkable results in the feature selection algorithm. However, the relevance analysis on features is not unique. In recent years, Ping et al. found a high level of label association between the label set and the feature set. Also, the labels are naturally clustered into several groups, indicating that similar labels tend to be clustered into the same group, and different labels belong to different groups. Based on this, they proposed a Multi-Label Feature Selection Considering the Max-Correlation in high-order label (MCMFS) [11] algorithm. The relevance analysis of features and label groups replaces the relevance analysis of features and labels. Although MCMFS only performs new relevance analysis without combining the traditional relevance analysis, but this work provides a new idea that learning the new correlations of features to do relevance analysis. For example, the treatment of dual correlation is recognized in causal feature selection [41].

In subsequent studies, it is found that the feature obtained from an instance is an external manifestation of the instance. When partial label information becomes experience, there should also be a correlation between the feature and the instance. For example, when people are ready to buy a vehicle, they will investigate the vehicle features within the label range after they know the label information such as the branding, model and applicable population. After people get enough information about the vehicle features in the label range, they will consider the vehicle's sample information to determine their purchasing needs. Therefore, learning from the data labels to obtain important label-specific feature information can effectively help the classification of the sample, and the label-instance information can help locate these important features. It is necessary to learn the correlation between label-specific features and label-instances.

A large number of experimental studies have shown that there is a connection between features and labels, and each label has the corresponding features [12-16]. According to the correlation of labels, Huang et al. proposed a Learning label specific features for multi-label classification algorithm (LLSF) that includes the correlation between labels [15]. Based on the label-specific-feature proposed by LLSF, a generic relationship matrix $\mathbf{W}$ that intuitively reflects the feature in the label can be obtained. Aiming at the generic relationship of the features in the labels, our study obtains the mutual information and the label

distribution information for relevance analysis. Then, the metric for a new relevance analysis was successfully obtained using the label-specific features extracted by the LLSF algorithm proposed by Huang et al. In this process, the correlation between features and labels is involved, so there will be a situation in which both correlations are considered. This means that the proposed algorithm needs two different relevance analysis modules, and the second relevance analysis module in series with the single channel structure will destroy the integrity of the first relevance analysis module thereby losing relevant features. Hence, the dual-channel feature selection method is proposed.

The dual-channel feature selection method is a parallel structure that independently solves the feature redundancy problem under the two different feature relevance analysis so as to alleviate the feature loss problem in the single-channel. Meanwhile, the related algorithms of the dual-channel structure are noticed [17,18]. Dual-channel is extremely rare in machine learning, but it is commonly used in deep learning to form dual-channel convolutional neural networks with CNN [19]. In recent years, dual-channel CNN has been applied to the research fields closely related to features, such as population feature engineering [20] and protein sequence feature fusion [21], indicating that the dual-channel structure can promote feature learning. These scientific research results provide certain theoretical support for the thought of this paper. Specifically, a dual-channel structure is applied to the relevant feature selection algorithms in the subspace, and a Parallel Dual-channel Multi-label Feature Selection (PDMFS) algorithm is proposed. The algorithm independently solves the feature redundancy under different feature relevance analysis by establishing parallel information channels and then merging the subsets to obtain the target feature set. This method attempts to select low-redundancy features under different feature correlations to obtain the best target feature set.

The main contributions of this paper are as follows:

1) Different from the current multi-label feature selection method, this paper utilizes the parallel dual-channel structure to successively solve the problem of feature redundancy under different correlations and the problem of subset feature redundancy.
2) Compared with the traditional multi-label feature selection, PDMFS consider both the correlation between features and labels and the correlation between label-specific features and label instances. PDMFS maximizes the original setting order of the pre-merged subsets by cross-merging methods to avoid feature confusion within the target feature set.
3) The experiment on eight benchmark multi-label datasets shows better performance of PDMFS against state-of-the-art multi-label feature selection algorithms in multi-label classification.

The rest of this paper is organized as follows. The second part introduces the related work of this paper. Then, the next part introduces the basic knowledge of theory. After that, presents details of proposed method in the fourth part. The comparison between the proposed algorithm and other advanced algorithms is shown in the next. The last part summarizes the full paper.

2. Related Work

As a common measurement metric, information entropy [22-25] has been widely used in feature selection. Lee et al. proposed a multi-label feature selection algorithm based on multi-variable mutual information to maximize the relevance between features and labels (PMU) [26]. Zhang et al. used two projection strategies to project the original data into a lower-dimensional feature space based on the maximum relevance between the original features and the labeled space. Then, they proposed an attribute reduction algorithm based on the maximum relevance (MDDMspc, MDDMproj) [27]. Recently, Amin et al. modeled the feature selection process as a multi-standard decision-making process for the first time, and they proposed multi-label feature selection with multi-standard decision making (MFS-MCDM) [28].

However, the above methods do not consider the redundancy between features. According to the mutual information between features and labels and the principle of maximizing the relevance and minimizing the redundancy [29], Liu et al. divided the feature local subspace [30] and performed a fixed ratio of feature selection, and they proposed a multi-label feature selection algorithm based on the local subspace (MFSLS) [31]. Based on the above research, Lin et al. proposed a multi-label feature selection algorithm based on neighborhood mutual information (MFNMIpes) [32]. They defined the neighborhood concept from different cognitive viewpoints and extended the neighborhood information entropy to multi-label learning.

In subsequent studies, MCMFS [11] replaces the feature-label correlation analysis by feature-label group correlation analysis through spectral clustering. Similarly, Aim et al. implemented a bi-objective optimal feature selection using Pareto-Clustering for feature relevance and redundancy, proposed an efficient Pareto-based feature selection algorithm for multi-label classification (PMFS) [36]. Regarding the feature selection algorithm combining label-specific features [37-40], Zhang et al. achieved full utilization of label information by learning label-specific features, proposed Fast multi-label feature selection via global relevance and redundancy optimization (GRROfast) [37] algorithm. In addition, Zhang et al. further utilized the label-specific features to present a new group-preserving label-specific feature selection (GLFS) [38] algorithm for multi-label learning, which simultaneously considers the features special to the labels in the same group and specific features owned by each label to execute feature selection. It can be found that multi-label feature selection algorithms using label-specific features for relevance analysis are gradually developed.

Although the above algorithms differ in the method of selecting feature subsets, they all follow the same rule: The study first determines the features relevant to the learning task and then conducts feature selection according to the requirements. For example, the MDDM, PMU, and MFS-MCDM algorithms all first obtain the feature ranking that satisfies the maximum relevance between features and labels, and then select an appropriate feature subset according to the requirements. These algorithms differ from the MFSLS, and MFNMIpes algorithms in the feature selection process. It is found that the feature with less redundancy is more in line with the needs of the learning task. Finally, a feature subset is obtained, which has the maximum relevance and the minimum redundancy [29]. Based on this, new feature relevance analysis such as label clustering and label-specific features have been combined, and multi-label feature selection algorithms such as MCMFS, PMFS, GRROfast and GLFS have been proposed successively. However, the relevance analysis is still singular.

3. Preliminaries

3.1. Information Entropy

By borrowing the concept of thermal entropy in thermodynamics, Shannon proposed information entropy in 1948. Information entropy describes the degree of uncertainty of the information, and it defines the amount of information in mathematical language. Let $A=\{a_1,a_2,...,a_n\}$ and $B=\{b_1,b_2,...,b_n\}$ be two discrete random variables. $P(a_i)$ is the probability of a_i, the information entropy [8,24,32] of A is

$$H(A)=-\sum_{i=1}^{n}P(a_i)log_2P(a_i). \tag{1}$$

The joint entropy [24] of A and B is

$$H(A,B)=-\sum_{i=1}^{n}\sum_{j=1}^{m}P(a_i,b_j)log_2P(a_i,b_j). \tag{2}$$

Given B, the conditional entropy [24] of A is

$$H(A\mid B)=H(A,B)-H(B). \tag{3}$$

When the information of random variable B is obtained, the entropy of random variable A is reduced, and the amount of reduction is the mutual information [24] of A and B

$$I(A;B)=H(A)-H(A\mid B). \tag{4}$$

$I(A;B)$ is used to measure the statistical correlation between A and B, and $I(A;B)\geq 0$. When $I(A;B)=0$, A and B are independent of each other, and no information is provided between the two variables.

As for multi-label feature selection, it is assumed that given two features x_1 and x_2 describing an instance, $I(x_1;x_2)$ can effectively describe the redundancy between x_1 and x_2; Given feature x of the description object and label set $Y=\{y_1,y_2,...,y_l\}$, $\forall y_i\in Y, i=1,2,...,l$, $I(x;y_i)$ can effectively describe the degree of correlation between the feature and the label. In this case, the mutual information between feature x and label set Y is

$$I(x;Y)=\sum_{i=1}^{l}I(x;y_i). \tag{5}$$

3.2. Subspace Feature Selection Method

In multi-label feature selection, the features with high redundancy are mostly irrelevant, but the features with a strong correlation may also contain high redundancy. This contradiction can be resolved by establishing a subspace model [30,31]. Based on this, the feature sequence that is sorted in descending order of the mutual information size between feature and label or between label-specific feature and label-instance can be divided into k subspaces, and then feature selection can be performed under the sampling ratio P. The specific process is as follows:

Given a feature space with a dimension of m, this space has k subspaces with a dimension of m/k, and the i-th space is $f_i=\{x_{i1},x_{i2},...,x_{i[m/k]}\}, \forall x_{ij}\in f_i$, where $j=1,2,...,[m/k]$. The mutual information between x_{ij} and other features is

$$R(x_{ij})=\sum_{\substack{q=1\\ q\neq j}}^{[m/k]}I(x_{ij},x_{iq}). \tag{6}$$

The smaller $R(x_{ij})$, the lower redundancy between x_{ij} and other features [29,31]. By ascendingly order the mutual information obtained by formula (6), a redundancy ranking of features $f_i'=\{x_{i1}',x_{i2}',...,x_{i[m/k]}'\}$ in the i-th subspace can be obtained.

Through the sampling ratio P, the feature with less redundancy in the i-th subspace is selected as a subset, and then the subsets are merged into the final subset.

3.3. Learning Label-specific Feature

In the LLSF proposed by Huang et al., it is assumed that each class label is associated with a feature subset from the original feature set. Compared with the original feature, the class label is sparse [15]. LLSF models the discriminative properties of label-specific features through linear regression, and it uses ℓ_1 -norms on regression parameters to model the sparsity of label-specific features.

Given the training dataset $D=\{d_1,d_2,...,d_n\}, \forall d_j\in D, j=1,2,...,n$. The feature set describing the data instance is $X=\{x_1,x_2,...,x_m\}, \forall x_t\in X, t=1,2,...,m$. Each instance may belong to a label set $Y=\{y_1,y_2,...,y_l\}, \forall y_i\in Y, i=1,2,...,l$. It is easy to obtain the feature matrix

$$\mathbf{X}=\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

and label matrix

$$\mathbf{Y}=\begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{l1} & \cdots & y_{ln} \end{bmatrix}.$$

Therefore, all the binary classifiers are considered while introducing label correlation in LLSF. The optimization model of label-specific features [15] can be expressed as:

$$\min_{W_i} \frac{1}{2}\|\mathbf{XW}-\mathbf{Y}\|_F^2+\frac{\alpha}{2}Tr(\mathbf{RW}^T\mathbf{W})+\beta\|\mathbf{W}\|_1 \tag{7}$$

where $\mathbf{W}=[\mathbf{W}_1,\mathbf{W}_2,...,\mathbf{W}_m]\in\mathbb{R}^{m\times l}$, $\mathbf{R}\in\mathbb{R}^{l\times l}$, $\alpha\geq 0$ and $\beta\geq 0$ are the model parameters.

Due to the non-smoothness of the ℓ_1 -norm regularization term, the objective function of the minimization problem of formula (7) is also non-smooth. Since this problem is a convex optimization problem, LLSF uses an accelerated near-end gradient method to solve this problem. The final matrix $\mathbf{W}$ is optimized as:

$$\mathbf{W}=\arg\min_{\mathbf{W}}\frac{1}{2}\left\|\mathbf{W}-G^{(t)}\right\|_F^2+\frac{\beta}{L_f}\|\mathbf{W}\|_1 \tag{8}$$

where $G^{(t)}=\mathbf{W}^{(t)}-\frac{1}{L_f}\nabla f(\mathbf{W}^{(t)})$, $f(\mathbf{W})=\frac{1}{2}\|\mathbf{XW}-\mathbf{Y}\|_F^2+\frac{\alpha}{2}Tr(\mathbf{RW}^T\mathbf{W})$, and L_f is Lipschitz constant. Based on this, the label-specific feature set $W=\{w_1,w_2,...,w_m\}$, $\forall w_t\in W$, $t=1,2,...,m$ can be obtained.

4. Proposed Method

4.1. Mutual Information between Label-specific Feature and Label-instance

Denote the mutual information between feature x_t and instance d_j as $I(x_t;d_j)$. In the label space, $d_j=\{y_{1j},y_{2j},...,y_{lj}\}$. Under the premise of knowing W, the correlation between features and instances is described with the mutual information between the label-specific feature and the label instance:

$$I(x_t;d_j)=I(w_t;d_j) \tag{9}$$

Then the correlation between features and instances can be defined as:

$$I(x_t;D)=\sum_{j=1}^{n}I(w_t;d_j)=I(w_t;D) \tag{10}$$

Similar to the characteristic of the mutual information between the feature and the label, the more the mutual information between the feature and the instance, the more important the feature is.

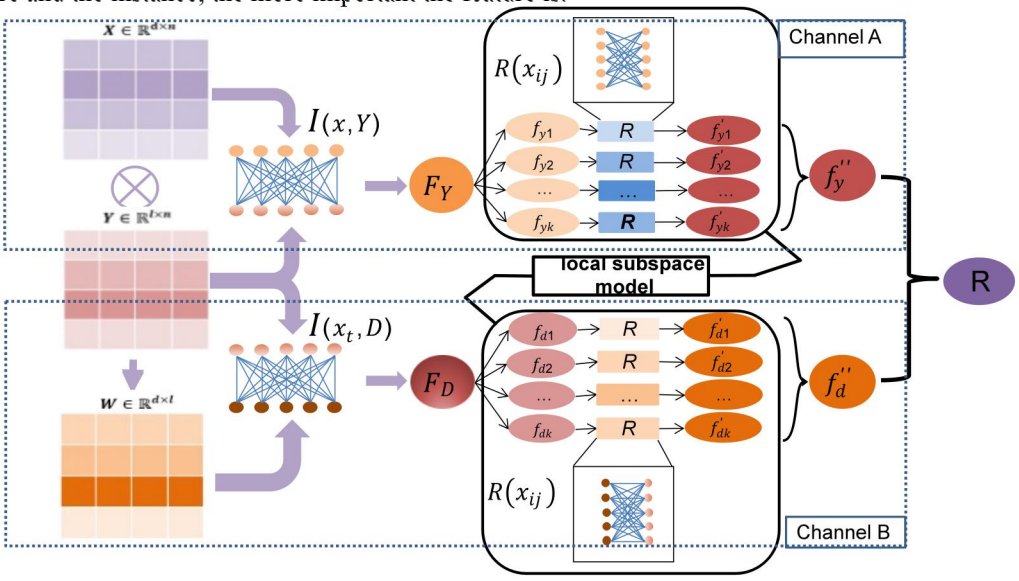

Figure 1. Flowchart of PDMFS

4.2. PDMFS: Parallel Dual-channel Multi-label Feature Selection

To obtain the features with maximizing the relevance and minimizing the redundancy between the feature and the label, the label-specific feature and the label-instance, this paper proposes a Parallel Dual-channel Multi-label Feature Selection algorithm (PDMFS). PDMFS designs the algorithm model as two independent modules, called Channel A and Channel B, and uses them to build a two-channel model. PDMFS performs subspace feature selection for correlation analysis between features and labels and between features and instances independently in the two channels. The specific process is shown in Figure 1.

From equations (5) and (10), it can be seen that the more mutual information between the feature and the label or the label-specific feature and the label instance, the more important the feature is. By descending the mutual information $I(x;Y)$ and $I(x_t;D)$ respectively, two sets of feature importance rankings F_Y and F_D can be obtained.

The two kinds of mutual information are sorted respectively by the subspace model in their respective channels, and the feature union subset obtained is the target feature subset

$$\mathrm{R} = f_y'' \cup f_d'', \tag{11}$$

where f_y'' and f_d'' represent the sum of the selected subspace feature subsets under the two sequences.

As shown in Figure 1, Channel A calculates the mutual information $I(x;Y)$ between features and labels as feature relevance analysis, and Channel B calculates the mutual information $I(x_t;D)$ between label-specific features obtained by LLSF and label instances as another feature relevance analysis. Subsequently, the feature importance rankings F_Y and F_D are independently performed for the maximum relevance minimum redundancy feature selection of the local subspace model. Finally, the channel feature subsets are merged to obtain the target feature subset.

According to the above description, the pseudocode of PDMFS is as follows:

Algorithm 1: Parallel Dual-channel Multi-label Feature Selection (PDMFS)

Input: Training dataset D , feature set X , label set Y ,Label-specific-feature set W , k is the number of subspaces divided and P_i is the sampling proportion of the *i-th* subspace in each subspace model,

$$\sum_{i}^{k} P_i = 1$$

Output: Feature subset R

(1) In channel A:

(2) $F_Y = \varnothing, f_y'' = \varnothing;$

(3) New feature subset f_y'' is obtained by subspace model;

(4) In channel B:

(5) for each $w_t \in W$

(6) $F_D = \varnothing, f_d'' = \varnothing$

(7) for each $y_i \in Y$

(8) Calculate $I(x;D)$ according to equation 10

(9) end

(10) end

(11) New feature sequence F_D is obtained by sorting $I(x;D)$ in descending order.

(12) The sequence F_D was evenly divided into k segments, and the i segment was represented by f_{di} ;

(13) for each f_{di}

(14) Calculate $R(x_{ij})$ according to equation 6 and sorted in ascending order respectively in channel B, a groups of new feature subsets f_d'' is obtained according to the proportion P_i ;

(15) end

(16) $\mathrm{R} = f_y'' \cup f_d''$

If PDMFS simply merging feature subsets, it will lead to the disorder of feature sequences in the original set. In this case, the target feature subset is internally disordered, resulting in poor performance stability of the algorithm. Therefore, this paper retains the original set order to the maximum extent by cross-merging method. It is known that the set of the subsets f_y'' and f_d'' can be obtained through parallel dual-channel, then the target feature set R acquisition can be modified as follows:

$$\left.\begin{aligned} f_y'' = \sum_{i=1}^{k} f_{yi}', f_d'' = \sum_{i=1}^{k} f_{di}', R_i = f_{yi}' \cup f_{di}', \\ \mathrm{R} = f_y'' \cup f_d'' = \sum_{i=1}^{k} f_{yi}' \cup f_{di}' = \sum_{i=1}^{k} R_i \end{aligned}\right\} \tag{12}$$

The specific process is shown in Figure 2.

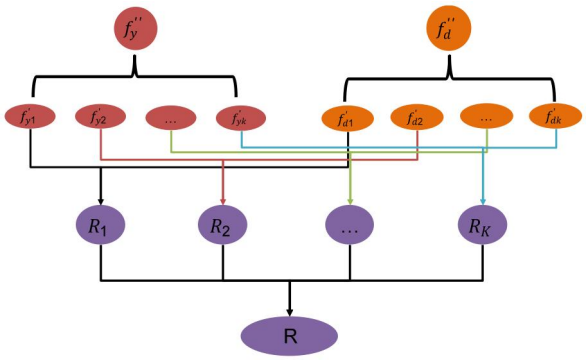

Figure 2. Flowchart of Cross-merging

As shown in Figure 2, Cross-merging first returns the feature subsets to the state before merging the local subspaces of the channel feature sets. Then the set of subspace features with the same level of relevance in each channel is merged. Finally, the merged subsets are summed to obtain the final subset in order to expect a more stable feature ranking within the feature subset.

4.3. Complexity Analysis

In PDMFS, suppose that the number of instances is n, the number of features is m, the number of labels is l, and the number of subspaces is k. The time complexity of the mutual information between the feature and the label is $O(ml)$. The time complexity of Channel A is $O(ml)$. The time complexity of the label-specific feature matrix is $O(m^2 + ml + l^2 + nd + nl)$, and the time complexity of the mutual information between the label-specific feature and the label-instance $O(mn)$. Thus, the time complexity of Channel B is $O(m^2 + ml + l^2 + nd + nl + mn)$ .The subspace redundancy feature selection time complexity is $O(m(m-k)/k)$. Suppose that the number of selected features is b, since the number of features chosen in this paper is the merging of two channel sets, the size range of b is $[(m/k),(2m/k)]$ .The time complexity of PMU, MFNMIpes, MCMFS and GRROfast are $O(mnl + bnml + nml^2)$, $O(n^2ml + (nm)^2 + bmn^2)$, $O(nml + bnl)$ and $O(Tvl + m^2 + ml)$ respectively, where T denotes the number of iterations for clustering and v is the number of groups (or cluster centers). Although PDMFS requires the calculation of two correlations, the monomial index of time complexity is much lower than PMU, MFNMIpes and MCMFS. Certainly, due to the group label search method of GRROfast, the time complexity of PDMFS is higher than that of GRROfast. And since the dual channels run in parallel, the total sum of time complexity of Channel A and Channel B does not produce exponential growth. Compared to the high exponential time complexity of PMU, MFNMIpes and MCMFS, the time complexity of PDMFS is still acceptable.

5. Experimental Data and Analysis of Results

5.1. Experimental Data

Table 1 Multi-label datasets

Datasets	Samples	Training	Test	Label	Feature	Domain
Scene[b]	1396	200	1196	6	294	Image
Business[b]	5000	2000	3000	30	438	Text
Computers[a]	5000	2000	3000	33	681	Text
Education[a]	5000	2000	3000	33	550	Text
Health[b]	5000	2000	3000	32	612	Text
Recreation[b]	5000	2000	3000	22	606	Text
Reference[b]	5000	2000	3000	33	793	Text
Science[a]	5000	2000	3000	40	743	Text

[a]Yahoo Web Pages (http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar).
[b]Mulan (http://mulan.sourceforge.net/datasets-mlc.html).

To verify the effectiveness of the proposed algorithms, eight datasets of Business, Computers, Education, Health, Recreation, Reference, Scene, and Science are used in the experiment. Table 1 shows detailed information about the eight multi-label datasets. Such as the type of domain from which the dataset was acquired; the numbers of samples, features, labels, training samples, and test samples, where the training and test samples are divided from the data source website.

5.2. Experimental Environment and Evaluation Index

The experimental codes are all implemented in Matlab2016a. The hardware platform is a computer equipped with Intel® Core(TM) i5-9600K CPU (3.70GHz) and 16 GB memory, and the computer runs Windows10 operating system. In this paper, six common multi-label learning evaluation indicators are taken to comprehensively evaluate the algorithm performance, including Average Precision (AP), Coverage (CV), Hamming Loss (HL), Ranking Loss (RL), Macro F1-score (Macro-F1), and Micro F1-score (Micro-F1) [33]. For convenience, the indicators are abbreviated as AP↑, CV↓, HL↓, RL↓, Macro-F1↑, and Micro-F1↑, where ↑ means the higher the value, the better the performance, and ↓ means the lower the value, the better the performance. Given the multi-label classifier $h(\cdot)$, the prediction function $f(\cdot,\cdot)$, the ranking function $rank_f$, and the multi-label dataset $D = \{(x_i, Y_i \mid 1 \le i \le n)\}$. The detailed calculation of the above six evaluation indicators is as follows:

1. AP: It evaluates the average score of the correct label permutation of a specific label. The value of this indicator is between 0 and 1, and the higher the value, the better the performance.

$$AP(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i|}\cdot\sum_{y\in Y_i}\frac{\left|\left\{rank_f(x_i,y^*) \le rank_f(x_i,y), y^* \in Y_i\right\}\right|}{rank_f(x_i,y)} \tag{13}$$

2. CV: It measures the steps it takes on average to traverse all relevant markers of the sample. The value of this indicator is greater than 0, and the lower the value, the better the performance.

$$CV(f) = \frac{1}{n}\sum_{i=1}^{n}\max_{y\in Y_i} rank_f(x_i,y) - 1 \tag{14}$$

3. HL: It measures the mismatch between the true label and the predicted label of the sample in the case of a single label. The value of this indicator is between 0 and 1, and the lower the value, the better the performance.

$$HL(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y|}\left|h(x_i) \ne Y_i\right| \tag{15}$$

4. RL: It considers the situation in which the ranking of the unrelated labels of the sample is lower than the ranking of the related labels. The value of this indicator is between 0 and 1, and the lower the value, the better the performance.

$$RL(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i|\left|\overline{Y_i}\right|}\cdot\left|\left\{(y_1,y_2) \mid f(x_i,y_1) \le f(x_i,y_2), (y_1,y_2) \in Y_i \times \overline{Y_i}\right\}\right| \tag{16}$$

F1-score: It measures the accuracy of a binary classification model in statistics. It takes into account both the accuracy rate and recall rate of the classification model. Its maximum value is 1, and the minimum value is 0. For the j-th label $y_i (1 \le j \le n)$, the two-class classification performance of the classifier $h(\cdot)$ on this label can be described by the following four statistics:

- TP_j (The number of "True" Positive examples)

$$TP_j = \left|\left\{x_i \mid y_i \in Y_i \wedge y_i \in h(x_i), (x_i, Y_i) \in D\right\}\right| \tag{17}$$

- FP_j (The number of "False" Positive examples)

$$FP_j = \left|\left\{x_i \mid y_i \notin Y_i \wedge y_i \in h(x_i), (x_i, Y_i) \in D\right\}\right| \tag{18}$$

- TN_j (The number of "True" Negative examples)

$$TN_j = \left|\left\{x_i \mid y_i \notin Y_i \wedge y_i \notin h(x_i), (x_i, Y_i) \in D\right\}\right| \tag{19}$$

- FN_j (The number of "False" Negative examples)

$$FN_j = \left|\left\{x_i \mid y_i \in Y_i \wedge y_i \notin h(x_i), (x_i, Y_i) \in D\right\}\right| \tag{20}$$

As for multi-class problems, Micro-F1 and Macro-F1 can be used for performance evaluation, with the maximum value of 1 and the minimum value of 0. Based on the above statistics:

5. Macro-F1: It is the arithmetic mean of the F1-scores of all labels. The value of this indicator is between 0 and 1, and the higher the value, the better the performance.

$$Macro-F1=\frac{1}{q}\sum_{i=1}^{q}\frac{2TP_i}{2TP_i+FP_i+TN_i} \tag{21}$$

6. Micro-F1: It can be seen as a weighted average of the F1 scores of all labels. The value of this indicator is between 0 and 1, and the higher the value, the better the performance.

$$Micro-F1=\frac{\sum_{i=1}^{q}2TP_i}{\sum_{i=1}^{q}2TP_i+FP_i+FN_i} \tag{22}$$

5.3. Parameter Settings and Experimental Results

The proposed algorithm PDMFS is compared with seven feature selection algorithms, namely Multi-Label Dimensionality Reduction via Dependence Maximization (MDDMspc, MDDMproj) [27], Feature Selection for Multi-label Classification Using Multivariate Mutual Information (PMU) [26], Multi-label Feature Selection algorithm based on Neighborhood Mutual Information (MFNMIpes) [31], Multi-label Feature Selection using Multi-Criteria Decision Making (MFS-MCDM) [28], Multi-Label Feature Selection Considering the Max-Correlation in high-order label (MCMFS) [11], Fast multi-label feature selection via Global Relevance and Redundancy Optimization (GRROfast) [37]. PMU, MDDMspc, MDDMproj and MFNMIpes as classical algorithms for information metrics in feature selection are used to compare the effectiveness of PDMFS performance. MFS-MCDM, MCMFS, and GRROfast as the state-of-the-art feature selection algorithms in recent years are used to compare the advancement of PDMFS. In the experiment, the δ parameter of the MDDMspc is set to 0.5; the label-specific feature coefficient matrix **W** of PDMFS is extracted by the LLSF algorithm, the matrix **W** is the post-five-fold cross-validation result and the α, β, and γ parameters of the LLSF algorithm range from $[2^{-10},2^{10}]$, $[2^{-10},2^{10}]$, and $\{0.1,1,10\}$; As the number of features in the dataset used for the experiments was $[294,793]$, it's not too large. In the literature [31], it has been experimentally demonstrated that the best prediction is achieved by dividing the subspace into three when the feature dimension is not too high. So, the number of subspaces k is set to 3 in PDMFS. About the sampling ratio P of the three subspaces, it has been experimentally demonstrated that the best prediction is achieved by set to $\{0.6,0.3,0.1\}$ in the literature [31]. For other algorithms taken for performance comparison, default parameter settings are used. Unlike wrapper methods and embedding methods, the filtered multi-label feature selection process is independent of the classifier. Therefore, the feature selection time loss and feature set performance evaluation of the PDMFS algorithm are not constrained by the classifier. In the experiments, to effectively compare the subset performance of feature selection of each algorithm, ML-KNN is used to evaluate the feature subset performance as a fair and effective classifier familiar in the field of multi-label feature selection. ML-KNN is a multi-label version of the KNN algorithm that the nearest neighbor number k is set to 10 and the smoothing coefficient is set to 1 [34].

Table 2. Average Precision performance for eight feature selection methods (↑).

Dataset	MDDMspc	MDDMproj	PMU	MFNMIpes	MFS-MCDM	MCMFS	GRROfast	PDMFS
Scene	0.6929	0.7109	0.7562	0.6640	0.7395	0.7488	0.7464	**0.7758**
Business	0.8687	0.8695	0.8627	0.8676	0.8655	0.8731	**0.8752**	0.8741
Computers	0.6253	0.6253	0.6278	0.6253	0.6230	0.6309	0.6347	**0.6428**
Education	0.5531	**0.5542**	0.5192	0.5072	0.5003	0.5269	0.5501	0.5368
Health	0.6680	0.6696	0.6612	0.6921	0.6412	**0.7131**	0.6881	0.7108
Recreation	0.4485	0.4456	0.4961	0.5007	0.4261	0.5215	0.5007	**0.5330**
Reference	0.6081	0.6118	0.6076	0.6156	0.5825	0.6256	**0.6558**	0.6314
Science	0.4477	0.4477	0.4472	0.4502	0.4048	0.4502	0.4693	**0.4735**
Average	5.4375	4.9375	5.7500	5.1250	7.5000	2.9375	2.5625	1.7500

The number of features selected by PDMFS is fixed, and the number of features obtained by other comparison algorithms is random. To better observe the changes in the indicators of each algorithm, all other algorithms use the same number of features as PDMFS. Table 2-Table 7 show the prediction performance of the eight algorithms MDDMspc, MDDMproj, PMU, MFNMIpes, MFS-MCDM, MCMFS, GRROfast, and PDMFS, where ***Average*** indicates the average ranking of each algorithm and the best experimental results are shown in bold.

From Tables 2-7, we can observe that PDMFS can obtain a better or comparable performance than any of the chosen comparison methods in the average ranking of all metrics. Although PDMFS still does not perform well on particular data, it does not affect the overall evaluation of the algorithm. For example, although the individual evaluation metrics are not optimal in the Reference dataset, most metrics are second only to state-of-the-art multi-label feature selection algorithms such as GRROfast, and

still significantly outperform classical algorithms such as PMU. In addition, PDMFS and other state-of-the-art algorithms performed weaker than classical algorithms in AP, CV and RL on the Education dataset. We realize that specific properties of the data can affect the performance of the algorithm. For example, when the number of truly relevant features in the feature distribution of the data is very sparse, the addition of other features can hardly offset the effect of redundancy even if those features are still extremely relevant. Comparing datasets Reference and Education with other datasets, it can be found that PDMFS performs well on most datasets, but considering the specific properties of the data, there is still room for improvement of PDMFS.

Furthermore, considering that the comparison algorithms are all multi-label feature selection algorithms for global search of features, while PDMFS is a multi-label feature selection algorithm for overall division of local search, the feature distribution of the dataset will have an impact on the results of both global and local methods. Comparing with the GRROfast algorithm, PDMFS outperformed in average ranking, which indicates that PDMFS achieves more correlation features using the dual-channel structure and proves the effectiveness of parallel dual relevance analysis.

Table 3. Coverage performance for eight feature selection methods (↓).

Dataset	MDDMspc	MDDMproj	PMU	MFNMIpes	MFS-MCDM	MCMFS	GRROfast	PDMFS
Scene	1.1873	1.0978	0.8846	1.2943	0.9548	0.9047	0.9356	**0.8010**
Business	2.3957	2.3733	2.4740	2.4427	2.4467	2.3720	2.3203	**2.3190**
Computers	4.4510	4.4510	4.5150	4.5450	4.4880	4.4020	4.3567	**4.3040**
Education	3.8577	**3.8507**	4.0993	4.2277	4.3267	4.1823	3.8963	3.9830
Health	3.3697	3.3583	3.5817	3.3450	3.7793	**3.2267**	3.2907	3.2310
Recreation	5.0180	5.0227	4.8467	4.8523	5.2883	4.7973	4.7850	**4.5125**
Reference	3.4660	3.4507	3.5380	3.5360	3.7807	3.4837	**3.0987**	3.4287
Science	6.9893	7.0023	7.0843	7.1960	7.4213	6.9647	**6.7170**	6.8013
Average	4.8125	4.4375	5.7500	6.3750	7.2500	3.3750	2.2500	1.7500

Table 4. Hamming loss performance for eight feature selection methods (↓).

Dataset	MDDMspc	MDDMproj	PMU	MFNMIpes	MFS-MCDM	MCMFS	GRROfast	PDMFS
Scene	0.1644	0.1572	0.1516	0.1679	0.1467	**0.1414**	0.1442	0.1452
Business	0.0283	0.0282	0.0286	0.0286	0.0284	0.0274	0.0275	**0.0272**
Computers	0.0411	0.0411	0.0402	0.0411	0.0406	0.0407	0.0406	**0.0399**
Education	0.0421	0.0421	0.0439	0.0440	0.0441	0.0431	0.0417	**0.0407**
Health	0.0449	0.0444	0.0467	0.0428	0.0488	**0.0418**	0.0428	0.0424
Recreation	0.0630	0.0632	0.0604	0.0600	0.0641	0.0584	0.0603	**0.0569**
Reference	0.0306	0.0305	0.0311	0.0315	0.0344	0.0292	**0.0280**	0.0308
Science	0.0350	0.0351	0.0344	0.0345	0.0354	0.0345	0.0346	**0.0342**
Average	5.5625	5.3125	5.0625	5.8125	6.6875	2.6875	3.0000	1.8750

Table 5. Ranking Loss performance for eight feature selection methods (↓).

Dataset	MDDMspc	MDDMproj	PMU	MFNMIpes	MFS-MCDM	MCMFS	GRROfast	PDMFS
Scene	0.2168	0.1976	0.1569	0.2357	0.1686	0.1603	0.1654	**0.1392**
Business	0.0437	0.0432	0.0460	0.0443	0.0446	0.0424	**0.0414**	**0.0414**
Computers	0.0925	0.0925	0.0942	0.0947	0.0949	0.0914	0.0901	**0.0894**
Education	0.0900	0.0900	0.0974	0.1007	0.1033	0.0989	**0.0914**	0.0941
Health	0.0634	0.0632	0.0686	0.0614	0.0729	**0.0583**	0.0611	0.0587
Recreation	0.1913	0.1919	0.1808	0.1799	0.1996	0.1756	0.1780	**0.1658**
Reference	0.0899	0.0895	0.0920	0.0919	0.0993	0.0898	**0.0788**	0.0882
Science	0.1394	0.1396	0.1421	0.1442	0.1500	0.1390	**0.1334**	0.1358
Average	4.8750	4.5000	5.7500	6.1250	7.5000	3.125	2.3125	1.8125

To verify whether the cross-merged subset in the model can improve the stability of the target feature set. We compared PDMFS with cross-merge and PDMFSN without cross-merge. Some results are as shown in Table 8, where *Metric* means indicator type. The variation of the indicator with the number of features selected is shown in Figure 3-10 and the interval for the number of features is five.

After cross-merging, the target feature set has improved in 70.83% of the results for the eight datasets and three experimental metrics. This demonstrates that without cross-merging, the target feature set is cluttered with internal features and its performance is not fully exploited. In the other 29.16% of results, the change was not significant and the majority of results had a difference between 0.0001 and 0.0006. Only the AP and CV metrics of the Reference dataset and the HL metrics of the Scene dataset showed relatively large performance changes, which were 0.0787, 0.0042, and 0.0084, respectively. Based on Figure 3 and Figure 9, we can see that the convergence of the metrics changes with cross-merging has improved compared to without cross-merging, and the overall performance tends to be stable. This indicates that the variation of feature sequences within the target feature set after cross-merging tends to be more convergent and stable in general, and the small performance loss is worth the cost.

Table 6. Macro-F1 performance for eight feature selection methods (↑).

Dataset	MDDMspc	MDDMproj	PMU	MFNMIpes	MFS-MCDM	MCMFS	GRROfast	PDMFS
Scene	0.2928	0.3383	0.3914	0.2823	0.4555	0.4964	**0.5368**	0.4588
Business	0.0521	0.0547	0.0417	0.0729	0.0622	0.0894	**0.1466**	0.0925
Computers	0.0803	0.0803	0.0767	0.0761	0.0547	0.0780	0.0717	**0.0878**
Education	0.0598	0.0614	0.0376	0.0292	0.0084	0.0399	**0.1249**	0.0646
Health	0.1506	0.1442	0.0969	0.0333	0.0752	0.1701	**0.2284**	0.1545
Recreation	0.0590	0.0559	0.0909	0.1114	0.0427	0.1269	0.1163	**0.1301**
Reference	0.0502	0.0535	0.0461	0.0579	0.0221	0.0573	**0.1463**	0.0715
Science	0.0396	0.0375	0.0332	0.0301	0.0102	0.0404	0.0507	**0.0598**
Average	5.0625	4.9375	6.0000	5.8750	7.0000	3.1250	2.1250	1.8750

Table 7. Micro-F1 performance for eight feature selection methods (↑).

Dataset	MDDMspc	MDDMproj	PMU	MFNMIpes	MFS-MCDM	MCMFS	GRROfast	PDMFS
Scene	0.3318	0.3510	0.4243	0.3095	0.5017	0.5056	**0.5418**	0.5030
Business	0.6768	0.6754	0.6721	0.6841	0.6746	0.6914	0.6877	**0.6951**
Computers	0.3870	0.3870	0.3657	0.3942	0.3534	0.4087	0.3840	**0.4453**
Education	0.1569	0.1439	0.0950	0.0596	0.0392	0.1108	0.1867	**0.1967**
Health	0.3776	0.3477	0.3595	0.0726	0.3488	**0.4938**	0.3998	0.4281
Recreation	0.0821	0.0835	0.1707	0.1908	0.0507	**0.2468**	0.1923	0.2393
Reference	0.3443	0.3561	0.3126	0.3596	0.1900	0.3421	**0.4307**	0.3890
Science	0.0904	0.0839	0.0931	0.0936	0.0274	0.1051	0.1188	**0.1376**
Average	5.1875	5.5625	6.0000	5.1250	7.1250	2.7500	2.6250	1.6250

Table 8. Part results of PDMFS and PDMFSN on 8 datasets.

Dataset	PDMFS	PDMFSN	PDMFS	PDMFSN	PDMFS	PDMFSN
Scene	**0.7758**	0.7604	**0.8010**	0.8219	0.1452	**0.1368**
Business	0.8741	**0.8744**	**2.3190**	2.3250	**0.0272**	**0.0272**
Computers	**0.6428**	0.6361	**4.3040**	4.3730	**0.0399**	0.0409
Education	0.5368	**0.5369**	**3.9830**	3.9907	**0.0407**	0.0416
Health	**0.7108**	0.7068	**3.2310**	3.2353	0.0424	**0.0421**
Recreation	**0.5330**	0.5040	**4.5125**	4.8190	**0.0569**	0.0592
Reference	0.6314	**0.6356**	3.4287	**3.3490**	0.0308	**0.0302**
Science	**0.4735**	0.4708	**6.8013**	6.8670	**0.0342**	**0.0342**
Metric	AP (↑)		CV (↓)		HL (↓)	

Most of the comparisons in Figure 3-Figure 10 support the conclusion that cross-merging enhances stability of the target feature set, but there are cases where the CV and HL metrics change abruptly and aggressively, as shown in Figure 6. So, there is a debate here, whether the abrupt variation in performance of feature selection represent unstable selection results? In most cases, yes. As abrupt changes in performance tend to bring about non-convergence of results, which is why this paper introduces cross-merging to obtain the target feature set. However, if we take the CV metric of the Education dataset as an example, we can see that the convergence in the last part of the graphical variation is significantly better than without cross-merging, although more drastic performance mutations occur after cross-merging. We are aware that the underlying structure of PDMFS is the subspace feature selection model and that the final target feature set is the sum of the subspace selection subsets. This means that there is inevitably a sudden change in performance in the subset-subset connection part. Therefore, the stability of the subspace feature selection target feature set can still be determined by the final convergence of it, and abrupt changes that do not affect the final convergence can be ignored.

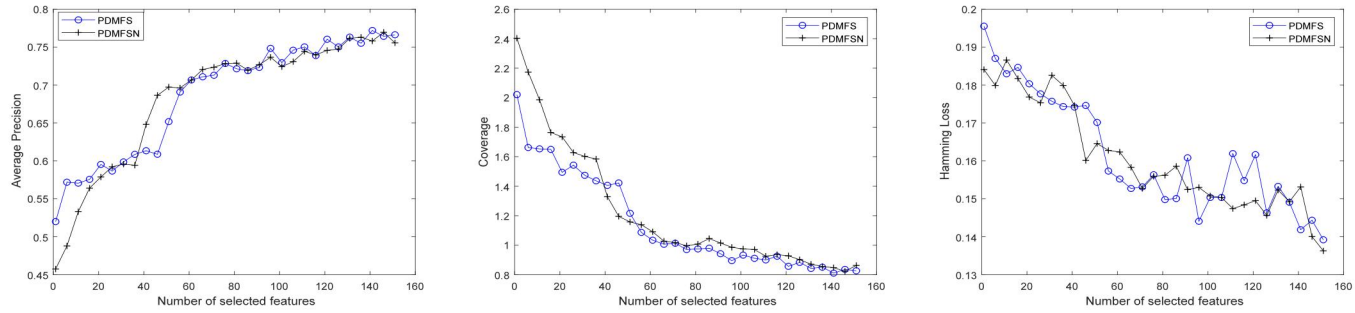

Figure 3. Experimental results on Scene dataset.

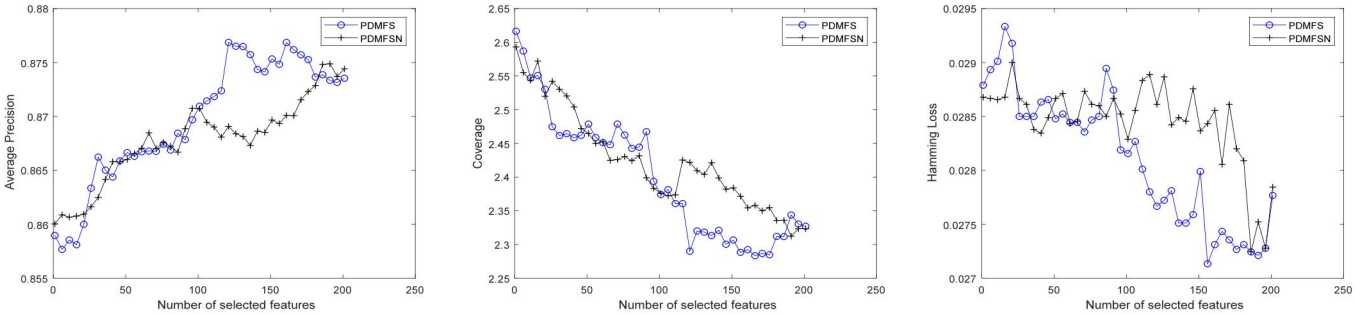

Figure 4. Experimental results on Business dataset.

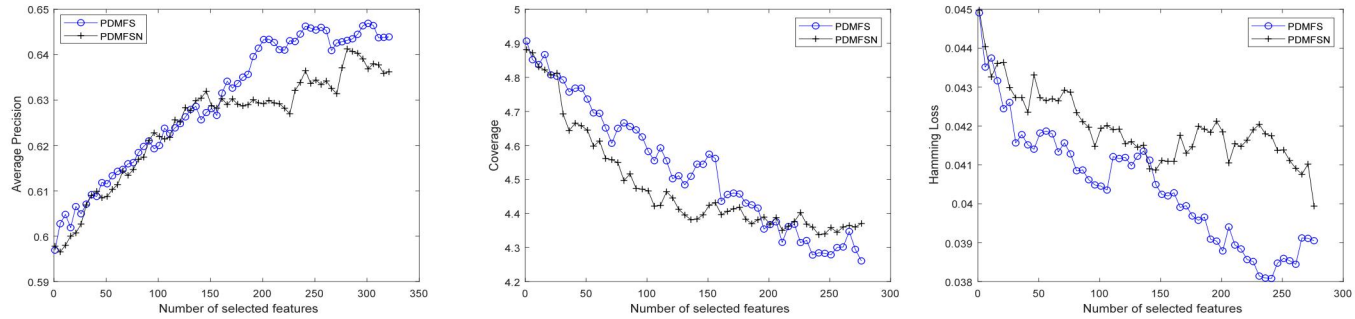

Figure 5. Part results on Computers dataset.

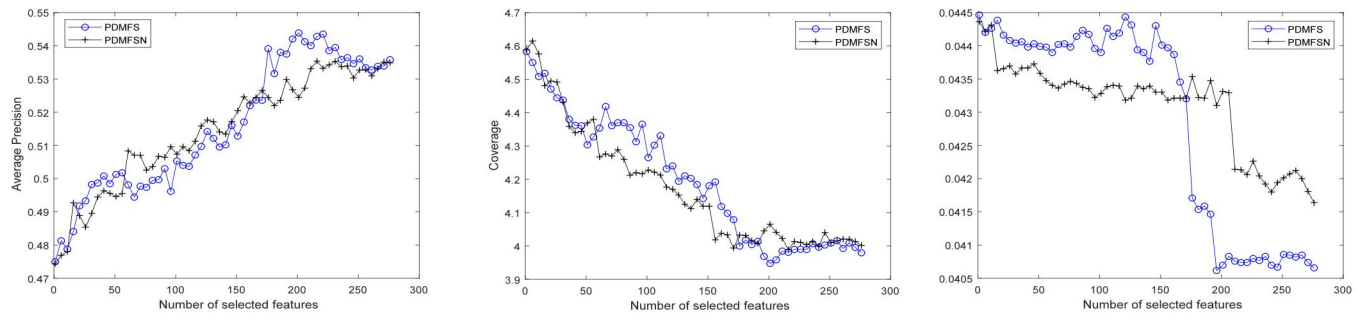

Figure 6. Part results on Education dataset.

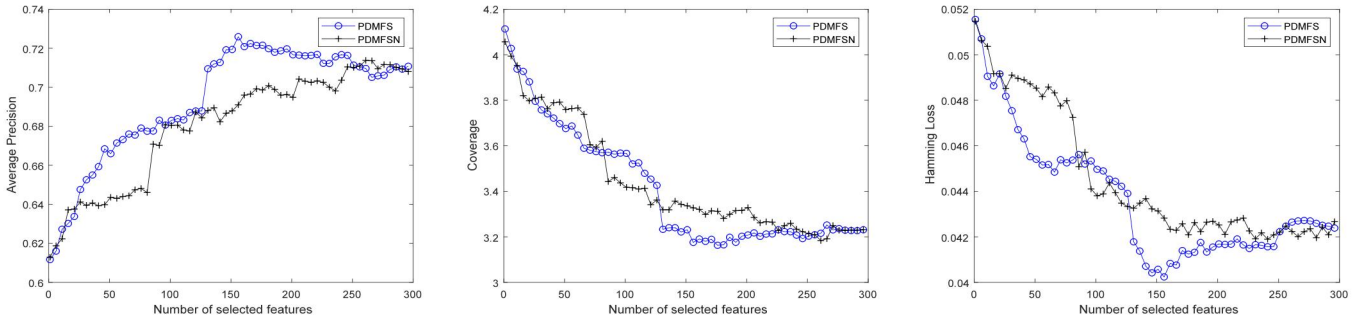

Figure 7. Part results on Health dataset.

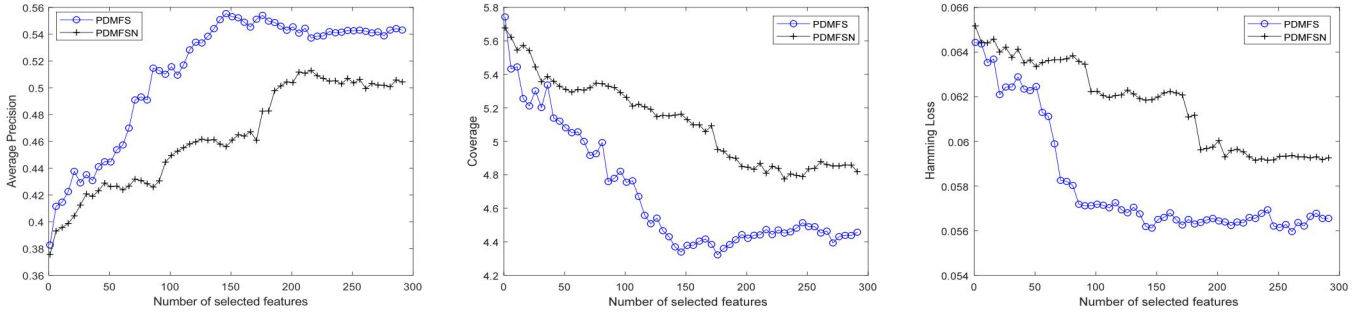

Figure 8. Part results on Recreation dataset.

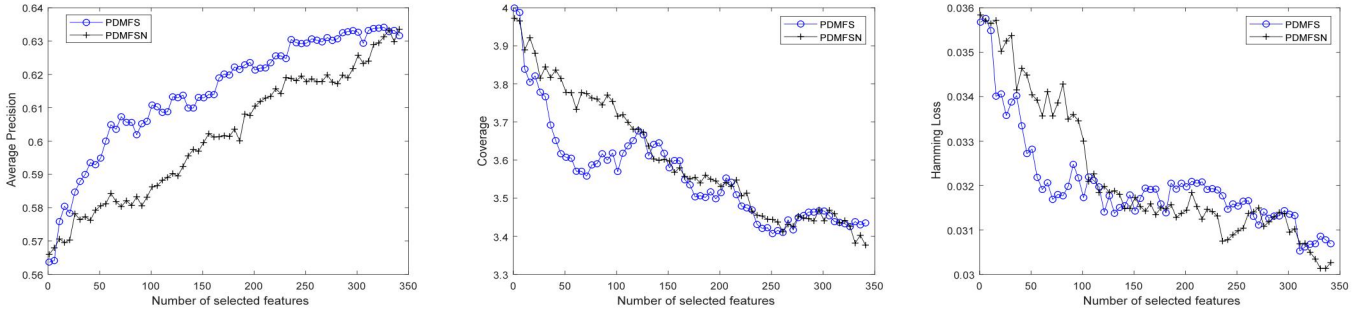

Figure 9. Part results on Reference dataset.

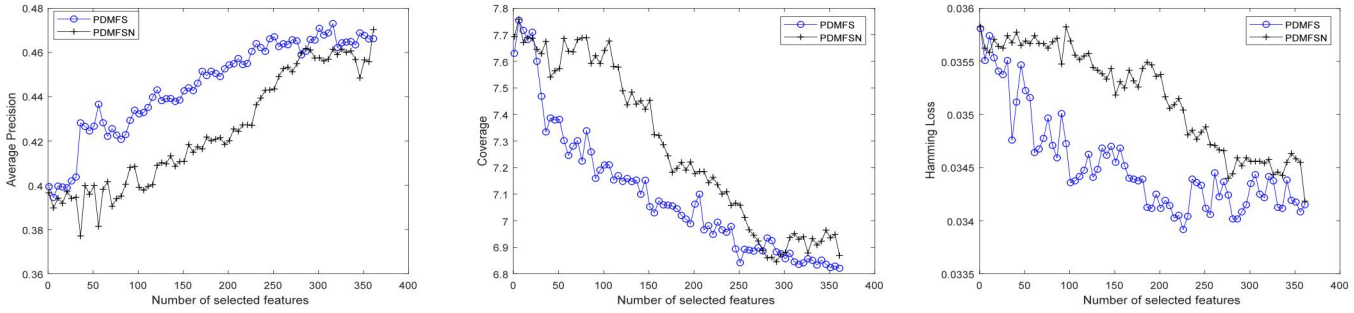

Figure 10. Part results on Science dataset.

5.4. Statistical Hypothesis Testing

This paper adopts the Nemenyi test under the significance level of $\alpha = 0.05$ [35] to evaluate the comprehensive performance of PDMFS and other algorithms. If the average ranking difference between the two comparison algorithms on all datasets is greater than the critical difference, the two algorithms are considered significantly different. Otherwise, there is no significant difference. The calculation of the CD value is as follows:

$$CD = q_\alpha \sqrt{\frac{K(K+1)}{6N}} \tag{23}$$

where $K = 8, N = 8, q_\alpha = 3.0310$, $CD = 3.7122$. Figure 11 shows the comparison between each algorithm on the six indicators of AP, CV, HL, RL, Macro-F1, and Micro-F1. The algorithms with no significant difference are connected by a solid line. The evaluation indicators are ordered from left to right, and the performance of the algorithm decreases accordingly.

For each algorithm, there are 42 kinds of experimental comparison results (7 comparison algorithms, 6 evaluation indicators). The following observations are obtained from the results shown in Figure 11. We can conclude that PDMFS ranks No.1 among all methods, and PDMFS is significantly different from the other algorithms in 54.77% of cases. Compared with the GRROfast algorithm, PDMFS fails to pull ahead, but under the condition of a specific feature subset size, PDMFS maintains the leading position. The results show that PDMFS has superior performance compared to the classical algorithms and good competitive performance compared to the state-of-the-art comparative algorithms.

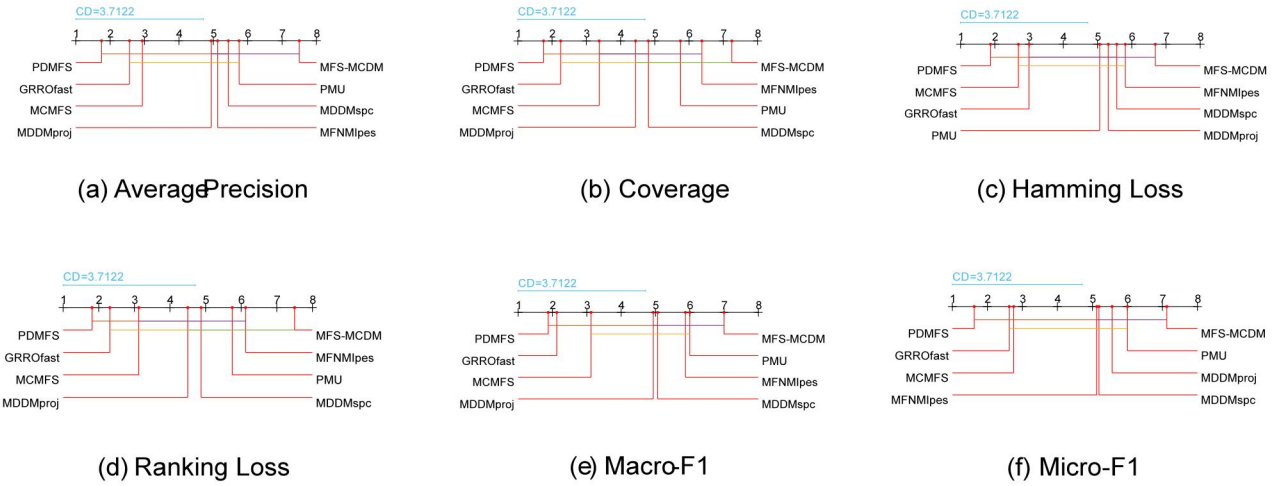

Figure 11. The performance comparison of algorithms

6. Conclusions

In this paper, a Parallel Dual-channel Multi-label Feature Selection algorithm is proposed, which explores a parallel structure to consider both the correlation between features and labels and the correlation between label-specific features and label instances. PDMFS adopts a minimal redundancy feature selection method under the dual correlation condition. Meanwhile, PDMFS solves the problem of losing important relevant features in single channel feature selection. The experimental results show that PDMFS includes more important and relevant features. On the other hand, cross-merging makes the target feature set of PDMFS more stable. However, PDMFS only obtains the final target feature set through the sum of subsets, ignoring the connections between feature subsets. From the subset perspective, each channel generates a subset of subspace features, and the minimum redundancy of the subset will be bound to be affected when the channel sets are merged. In other words, the redundancy of the target feature set obtained by PDMFS cross-merging is increased compared to the pre-merge subset. This suggests that the selection model of PDMFS needs to consider the uniformity of properties between sets, especially subsets and merged sets. This will be a major direction for our future research.

Conflict of interest: The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgments: This work was supported by the National Natural Science Foundation of Anhui under Grant 2108085MF216; the Key Laboratory of Data Science and Intelligence Application, Fujian Province University (NO. D202005); and the Graduate Academic Innovation Program of Anqing Normal University.

References

1. Zhang ML, Zhou ZH (2013) A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering 26(8): 1819–1837
2. Zhang P, Liu G, Gao W, et al (2021) Multi-label feature selection considering label supplementation. Pattern Recognition 120: 108137
3. Zhang L, Hu Q, Duan J, et al (2014) Multi-label feature selection with fuzzy rough sets. In *International Conference on Rough Sets and Knowledge Technology*; Springer International Publishing: Cham, Switzerland, pp 121–128
4. Spolaôr N, Cherman EA, Monard MC, et al (2013) A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science 292: 135–151
5. Fan Y, Liu J, Weng W, et al (2021) Multi-label feature selection with constraint regression and adaptive spectral graph. Knowledge-Based Systems 212: 106621
6. Huang R, Wu Z (2021) Multi-label feature selection via manifold regularization and dependence maximization. Pattern Recognition 120: 108149
7. Fan Y, Liu J, Weng W, et al (2021) Multi-label feature selection with local discriminant model and label correlations. Neurocomputing 442: 98-115
8. Li Y, Cheng Y (2019) Streaming Feature Selection for Multi-Label Data with Dynamic Sliding Windows and Feature Repulsion Loss. Entropy 21(12): 1151

9. Hu L, Li Y, Gao W, et al (2020) Multi-label feature selection with shared common mode. Pattern Recognition 104: 107344
10. Jiang L, Yu G, Guo M, et al (2020) Feature selection with missing labels based on label compression and local feature correlation. Neurocomputing 395: 95-106
11. Zhang P, Gao W, Hu J, et al (2020) Multi-label feature selection based on high-order label correlation assumption. Entropy 22(7): 797
12. Wang Y, Zheng W, Cheng Y, et al (2020) Joint label completion and label-specific features for multi-label learning algorithm. Soft Computing 24(9): 6553–6569
13. Cheng YS, Zhang C, Pang SF (2022) Multi-label space reshape for semantic-rich label-specific features learning. International Journal of Machine Learning and Cybernetics 13(4):1005–1019
14. Wang Y, Zheng W, Cheng Y, et al (2021) Two-level label recovery-based label embedding for multi-label classification with missing labels. Applied Soft Computing 99: 106868
15. Huang J, Li G, Huang Q, et al (2015) Learning label specific features for multi-label classification. Proceedings of the IEEE International Conference on Data Mining. IEEE, Atlantic City, New Jersey, USA pp: 181-190
16. Zhang L, Cheng T, Wang Y, et al (2021) Feature-label dual-mapping for missing label-specific features learning. Soft Computing 25(14): 9307–9323
17. Cui X, Zou C, Wang Z (2021) Remote sensing image recognition based on dual-channel deep learning network. Multimedia Tools and Applications 80(18): 27683–27699
18. Li H, Zheng Y, Ren P (2019) Dual-channel attention model for text sentiment analysis. International Journal of Performability Engineering. 15(3): 834-841
19. Zhou F, Ma Y, Wang B, et al (2021) Dual-channel convolutional neural network for power edge image recognition. Journal of Cloud Computing 10(1): 1-9
20. Wang X, Liu Y, Du Z, et al (2021). Prediction of protein solubility based on sequence feature fusion and DDcCNN. Interdisciplinary sciences: computational life sciences 13(4): 703–716
21. Xu Y, Lu L, Xu Z, et al (2019) Dual-channel CNN for efficient abnormal behavior identification through crowd feature engineering. Machine Vision and Applications 30(5): 945–958
22. Lee J, Kim DW (2015) Memetic feature selection algorithm for multi-label classification. Information Sciences 293: 80-96
23. Lee J, Kim DW (2015) Mutual information based multi-label feature selection using interaction information. Expert Systems with Applications 42(4): 2013-2025
24. Lin Y, Hu X, Wu X (2014) Quality of information-based source assessment and selection. Neurocomputing 133: 95-102
25. Estrela G, Gubitoso MD, Ferreira CE, et al (2020) An Efficient, Parallelized Algorithm for Optimal Conditional Entropy-Based Feature Selection. Entropy 22(4): 492
26. Lee J, Kim DW (2013) Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters 34(3): 349-357
27. Zhang Y, Zhou Z H (2010) Multilabel dimensionality reduction via dependence maximization. ACM Transactions on Knowledge Discovery from Data 4(3): 1-21
28. Amin H, Mohammad BD, Hossein N (2020) MFS-MCDM: Multi-label feature selection using multi-criteria decision making. Knowledge-Based Systems 206: 106365
29. Lin Y, Hu Q, Liu J (2015) Multi-label feature selection based on max-dependency and min-redundancy. Neuro Computing 168: 92-103
30. Zeng Z, Wang X, Chen Y (2017) Multimedia annotation via semi-supervised shared-subspace feature selection. Journal of Visual Communication and Image Representation 48: 386-395
31. Liu J, Lin M, Wang C, et al (2016) Multi-label feature selection algorithm based on local subspace. Pattern Recognition and Artificial Intelligence 29(3): 240-251
32. Lin Y, Hu Q, Liu J, et al (2016) Multi-label feature selection based on neighborhood mutual information. Applied Soft Computing 38: 244-256
33. Schapire RE, Singer Y (2000) BoosTexter: A Boosting-based System for Text Categorization. *Mach. Learn* 39(2): 135–168
34. Zhang ML, Zhou ZH (2007) ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40(7): 2038-2048
35. Janez D, Dale S (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7: 1-30
36. Aim H, Mohammad DB, Hossein N (2021) An efficient Pareto-based feature selection algorithm for multi-label classification. Information Sciences 581: 428-447
37. Zhang J, Lin Y, Jiang M, et al (2022) Fast Multilabel Feature Selection via Global Relevance and Redundancy Optimization. IEEE Transactions on Neural Networks and Learning Systems: 1-14, Doi: 10.1109/TNNLS.2022.3208956
38. Zhang J, Wu H, Jiang M, et al (2023) Group-preserving label-specific feature selection for multi-label learning. Expert Systems with Applications 213: 118861
39. Yu K, Cai M, Wu X, et al (2021) Multilabel feature selection: a local causal structure learning approach. IEEE Transactions on Neural Networks and Learning Systems: 1-14, Doi: 10.1109/TNNLS.2021.3111288
40. Wu X, Jiang B, Yu K, et al (2020) Multi-label causal feature selection. Proceedings of the AAAI Conference on Artificial Intelligence 34(04): 6430-6437
41. Guo X, Yu K, Liu L, et al (2022) Causal Feature Selection with Dual Correction. IEEE Transactions on Neural Networks and Learning Systems: 1-14, Doi: 10.1109/TNNLS.2022.3178075