

Parallel Dual-channel Multi-label Feature Selection

Jiali Miao

Anqing Normal University

Yibin Wang

Anqing Normal University

Yusheng CHENG (✉ chengyshaq@163.com)

Anqing Normal University <https://orcid.org/0000-0002-6562-1153>

Fei Chen

Tongling University

Research Article

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

Posted Date: September 2nd, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1436519/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Parallel Dual-channel Multi-label Feature Selection

Jiali Miao¹, Yibin Wang^{1,2}, Yusheng Cheng^{1,2,*}, Fei Chen³

¹ School of Computer and Information, Anqing Normal University, Anhui, Anqing, 246011, China; 1150838286@qq.com (J.M.); wbyjqwy@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 serial structures 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 dual-channel concept and design the algorithm model as two independent modules obtaining different feature correlation sequences, so that the algorithm can avoid the over-reliance on single feature correlation. 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; the wrapper methods use the accuracy of a particular classifier to determine the selected feature quality; and the embedding methods 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.

If the feature correlation and selection requirements are regarded as two modules, the current algorithms serial arrange the two modules in normal. In this study, this type of structure that deals with single correlation and selection requirements in series is called a single-channel structure. This structure has achieved remarkable results in the feature selection algorithm based on the correlation between features and labels. In fact, the correlation learning of features is not unique. In recent years, Ping et al. found that there are high-level label correlations in the label 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]. This provides a new approach to learn the new correlations of features. 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 a corresponding feature [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 correlation analysis. Then, the metric for a new correlation 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. At this time, there are two choices. One is to use a single-channel structure to continuously process the feature redundancy under the two correlations. However, the important related features will be lost, so this method is no longer applicable. The other choice is to use our proposed dual-channel feature selection method.

The dual-channel feature selection method is a parallel structure that independently solves the feature redundancy problem under the two correlations 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 algorithm is proposed. The algorithm independently solves the feature redundancy under different feature correlations 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 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 correlation 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 correlation between the original features and the labeled space. Then, they proposed an attribute reduction algorithm based on the maximum correlation (MDDM_{spc}, MDDM_{proj}) [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 correlation 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.

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 a feature ranking that satisfies the greatest correlation between features and labels, and then they 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 correlation between the feature and the label and the minimum redundancy between the feature and the feature [29].

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 $A = \{a_1, a_2, \dots, a_n\}$ and $B = \{b_1, b_2, \dots, b_m\}$ be two discrete random variables. $P(a_i)$ is the probability of a_i , the information entropy [8,24,32] of A is

$$H(A) = -\sum_{i=1}^n P(a_i) \log_2 P(a_i). \quad (1)$$

The joint entropy [24] of A and B is

$$H(A, B) = -\sum_{i=1}^n \sum_{j=1}^m P(a_i, b_j) \log_2 P(a_i, b_j). \quad (2)$$

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

$$H(A|B) = H(A, B) - H(B). \quad (3)$$

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

$$I(A; B) = H(A) - H(A|B). \quad (4)$$

$I(A; B)$ 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 $Y = \{y_1, y_2, \dots, y_l\}$,

$\forall y_i \in Y, i=1,2,\dots,l, I(x; y_i)$ 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:

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

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 $f_i = \{x_{i1}, x_{i2}, \dots, x_{i[m/k]}\}, \forall x_{ij} \in f_i$, where $j=1,2,\dots,[m/k]$. The mutual information between x_{ij} and other features is

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

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 $f_i^t = \{x'_{i1}, x'_{i2}, \dots, x'_{i[m/k]}\}$ 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 $D = \{d_1, d_2, \dots, d_n\}, \forall d_j \in D, j=1,2,\dots,n$. The feature set describing the data instance is $X = \{x_1, x_2, \dots, x_m\}, \forall x_t \in X, t=1,2,\dots,m$. Each instance may belong to a label set $Y = \{y_1, y_2, \dots, y_l\}, \forall y_i \in Y, i=1,2,\dots,l$. It is easy to obtain the feature matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \mathbf{M} & \mathbf{O} & \mathbf{M} \\ x_{m1} & \dots & x_{mn} \end{bmatrix}$$

and label matrix

$$\mathbf{Y} = \begin{bmatrix} y_{11} & \dots & y_{1n} \\ \mathbf{M} & \mathbf{O} & \mathbf{M} \\ y_{l1} & \dots & y_{ln} \end{bmatrix}.$$

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:

$$\min_{\mathbf{W}_i} \frac{1}{2} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_F^2 + \frac{\alpha}{2} \text{Tr}(\mathbf{R}\mathbf{W}^T \mathbf{W}) + \beta \|\mathbf{W}\|_1. \quad (7)$$

where $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_m] \in \mathbf{R}^{m \times l}, \mathbf{R} \in \mathbf{R}^{l \times l}, \alpha \geq 0$ 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:

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

where $G^{(t)} = \mathbf{W}^{(t)} - \frac{1}{L_f} \nabla f(\mathbf{W}^{(t)})$, $f(\mathbf{W}) = \frac{1}{2} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_F^2 + \frac{\alpha}{2} \text{Tr}(\mathbf{R}\mathbf{W}^T\mathbf{W})$, and L_f is Lipschitz constant. Based on this, the label-specific feature set $W = \{w_1, w_2, \dots, w_m\}$, $\forall w_t \in W, t = 1, 2, \dots, m$ 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, $d_j = \{y_{1j}, y_{2j}, \dots, y_{lj}\}$. 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:

$$I(x_t; d_j) = I(w_t; d_j) \quad (9)$$

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

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

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.

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

To obtain the features with the greatest correlation and the least 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.

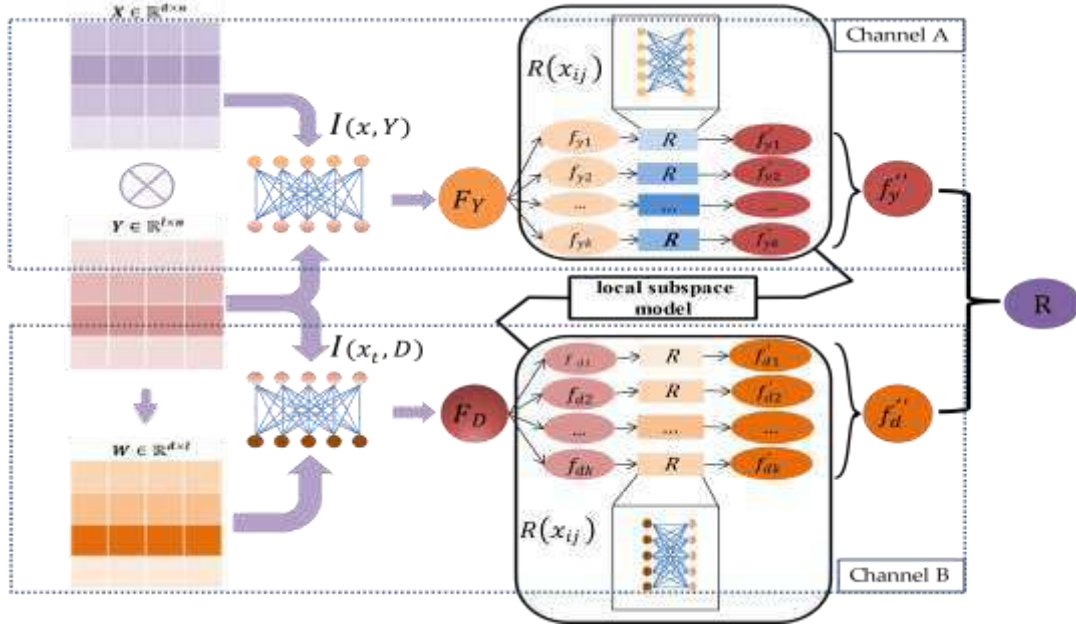


Figure 1. Flowchart of PDMFS

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

$$R = f'_y \cup f'_d, \quad (11)$$

where f'_y and f'_d represent the sum of the selected subspace feature subsets under the two sequences. 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) $F_Y = \emptyset, f_y'' = \emptyset$;
 - (3) New feature subset f_y'' is obtained by subspace model;
 - (4) In channel B:
 - (5) for each $w_i \in W$
 - (6) $F_D = \emptyset, f_d'' = \emptyset$;
 - (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) $R = f_y'' \cup f_d''$
-

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 retain 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:

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

The specific process is shown in Figure 2.

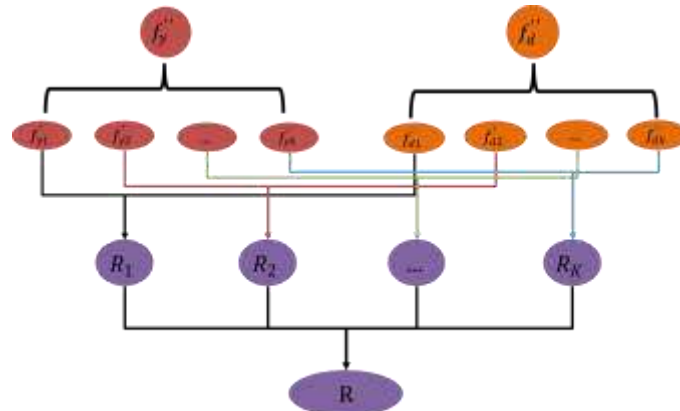


Figure 2. Flowchart of cross-merging

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 subspace is k . The time complexity of the mutual information between the feature and the label is $O(ml)$. So 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 $O(mn)$. So 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 and MCMFS respectively is $O(mnl + bnml + nml^2)$, $O(n^2ml + (nm)^2 + bmn^2)$ and $O(nml + bnl)$. Although PDMFS requires the calculation of two correlations, the monomial index of time complexity is much lower than PMU, MFNMIpes and MCMFS. 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

To verify the effectiveness of the proposed algorithms, ten 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.

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

^aYahoo Web Pages (<http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar>).

^bMulan (<http://mulan.sourceforge.net/datasets-mlc.html>).

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 \uparrow , CV \downarrow , HL \downarrow , RL \downarrow , Macro-F1 \uparrow , and Micro-F1 \uparrow , where \uparrow means the higher the value, the better the performance, and \downarrow 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

$D = \{(x_i, Y_i | 1 \leq i \leq n)\}$. 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 maximum value of this indicator is 1, and the minimum value is 0.

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

2. CV: It measures the steps it takes on average to traverse all relevant markers of the sample.

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

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 maximum value of this indicator is 1, and the minimum value is 0.

$$HL(h) = \frac{1}{n} \sum_{i=1}^n \frac{1}{|Y|} |h(x_i) \neq Y_i| \quad (15)$$

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 maximum value of this indicator is 1, and the minimum value is 0.

$$RL(f) = \frac{1}{n} \sum_{i=1}^n \frac{1}{|Y_i| |\bar{Y}_i|} \cdot \left| \left\{ (y_1, y_2) \mid f(x_i, y_1) \leq f(x_i, y_2), (y_1, y_2) \in Y_i \times \bar{Y}_i \right\} \right| \quad (16)$$

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 ($1 \leq j \leq n$), 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)

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

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

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

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

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

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

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

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

5. Macro-F1: It is the arithmetic mean of the F1-scores of all labels.

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

6. Micro-F1: It can be seen as a weighted average of the F1 scores of all labels.

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

5.3. Parameter Settings and Experimental Results

The proposed algorithm PDMFS is compared with five feature selection algorithms, namely Multi-Label Dimensionality Reduction via Dependence Maximization (MDDM_{spc}, MDDM_{proj}) [27], Feature Selection for Multi-label Classification Using Multivariate Mutual Information (PMU) [26], Multi-label Feature Selection algorithm based on Neighborhood Mutual Information (MFNM_{Ipes}) [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]. PMU, MDDM_{spc}, MDDM_{proj}, MFNM_{Ipes} and as classical algorithms for information metrics in feature selection are used to compare the effectiveness of PDMFS performance. MFS-MCDM and MCMFS as the latest feature selection algorithms in recent years are used to compare the advancement of PDMFS. In the experiment, the δ parameter of the MDDM_{spc} is set to 0.5; the label-specific feature coefficient matrix \mathbf{W} of PDMFS is extracted by the LLSF algorithm, the matrix \mathbf{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 [30], 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. In the experiment, kNN is used as the classifier to verify the performance of the algorithms, and the parameters of

ML-kNN are set to the default (i.e., the smoothing coefficient is set to 1, and the nearest neighbor number k is set to 10) [34]. 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 seven algorithms MDDM_{spc}, MDDM_{proj}, PMU, MFNMI_{pes}, MFS-MCDM, MCMFS and PDMFS, where *Average* indicates the average ranking of each algorithm and the best experimental results are shown in bold.

Table 2. Average Precision performance for seven feature selection methods (\uparrow).

Dataset	MDDM _{spc}	MDDM _{proj}	PMU	MFNMI _{pes}	MFS-MCDM	MCMFS	PDMFS
Scene	0.6929	0.7109	0.7562	0.6640	0.7395	0.7488	0.7758
Business	0.8687	0.8695	0.8627	0.8676	0.8655	0.8731	0.8741
Computers	0.6253	0.6253	0.6278	0.6253	0.6230	0.6309	0.6428
Education	0.5531	0.5542	0.5192	0.5072	0.5003	0.5269	0.5368
Health	0.6680	0.6696	0.6612	0.6921	0.6412	0.7131	0.7108
Recreation	0.4485	0.4456	0.4961	0.5007	0.4261	0.5215	0.5330
Reference	0.6081	0.6118	0.6076	0.6156	0.5825	0.6256	0.6314
Science	0.4477	0.4477	0.4472	0.4502	0.4048	0.4502	0.4735
<i>Average</i>	4.5625	4.0625	4.8750	4.3125	6.5000	2.3125	1.3750

Table 3. Coverage performance for seven feature selection methods (\downarrow).

Dataset	MDDM _{spc}	MDDM _{proj}	PMU	MFNMI _{pes}	MFS-MCDM	MCMFS	PDMFS
Scene	1.1873	1.0978	0.8846	1.2943	0.9548	0.9047	0.8010
Business	2.3957	2.3733	2.4740	2.4427	2.4467	2.3720	2.3190
Computers	4.4510	4.4510	4.5150	4.5450	4.4880	4.4020	4.3040
Education	3.8577	3.8507	4.0993	4.2277	4.3267	4.1823	3.9830
Health	3.3697	3.3583	3.5817	3.3450	3.7793	3.2267	3.2310
Recreation	5.0180	5.0227	4.8467	4.8523	5.2883	4.7973	4.5125
Reference	3.4660	3.4507	3.5380	3.5360	3.7807	3.4837	3.4287
Science	6.9893	7.0023	7.0843	7.1960	7.4213	6.9647	6.8013
<i>Average</i>	3.9375	3.5625	4.8750	5.3750	6.2500	2.6250	1.3750

Table 4. Hamming loss performance for seven feature selection methods (\downarrow).

Dataset	MDDM _{spc}	MDDM _{proj}	PMU	MFNMI _{pes}	MFS-MCDM	MCMFS	PDMFS
Scene	0.1644	0.1572	0.1516	0.1679	0.1467	0.1414	0.1452
Business	0.0283	0.0282	0.0286	0.0286	0.0284	0.0274	0.0272
Computers	0.0411	0.0411	0.0402	0.0411	0.0406	0.0407	0.0399
Education	0.0421	0.0421	0.0439	0.0440	0.0441	0.0431	0.0407
Health	0.0449	0.0444	0.0467	0.0428	0.0488	0.0418	0.0424
Recreation	0.0630	0.0632	0.0604	0.0600	0.0641	0.0584	0.0569
Reference	0.0306	0.0305	0.0311	0.0315	0.0344	0.0292	0.0308
Science	0.0350	0.0351	0.0344	0.0345	0.0354	0.0345	0.0342
<i>Average</i>	4.5625	4.3125	4.3125	5.1250	5.7500	2.3125	1.6250

The comparison between the proposed algorithms and other algorithms is as follows:

1. In terms of average precision, PDMFS achieves the best performance on six datasets. In the other datasets, the Education dataset where PDMFS performed the worst was only 0.0174 lower than the best comparison algorithm MDDM_{proj}. PDMFS still ranked the top three and outperformed most algorithms.
2. For the coverage indicator listed in Table 3, PDMFS achieves the best coverage on seven datasets.

3. In terms of the Hamming loss in Table 3, PDMFS achieves the best results on five datasets. According to the average ranking results of the eight datasets, PDMFS performs the best.

Table 5. Ranking Loss performance for seven feature selection methods (\downarrow).

Dataset	MDDM _{spc}	MDDM _{proj}	PMU	MFNMI _{pes}	MFS-MCDM	MCMFS	PDMFS
Scene	0.2168	0.1976	0.1569	0.2357	0.1686	0.1603	0.1392
Business	0.0437	0.0432	0.0460	0.0443	0.0446	0.0424	0.0414
Computers	0.0925	0.0925	0.0942	0.0947	0.0949	0.0914	0.0894
Education	0.0900	0.0900	0.0974	0.1007	0.1033	0.0989	0.0941
Health	0.0634	0.0632	0.0686	0.0614	0.0729	0.0583	0.0587
Recreation	0.1913	0.1919	0.1808	0.1799	0.1996	0.1756	0.1658
Reference	0.0899	0.0895	0.0920	0.0919	0.0993	0.0898	0.0882
Science	0.1394	0.1396	0.1421	0.1442	0.1500	0.1390	0.1358
<i>Average</i>	4.0000	3.6250	4.8750	5.1250	6.5000	2.5000	1.3750

Table 6. Macro-F1 performance for seven feature selection methods (\uparrow).

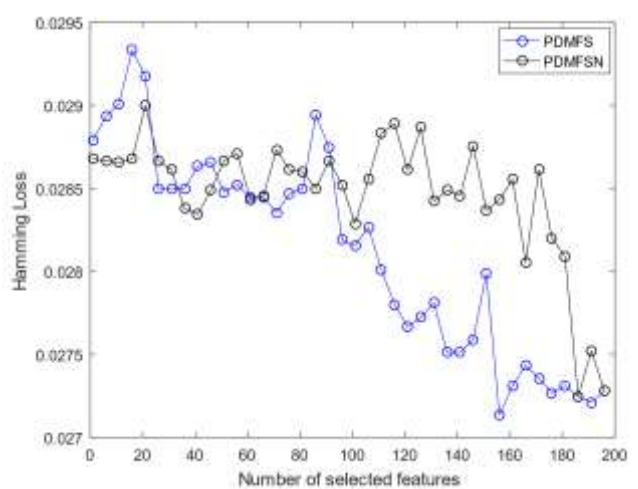
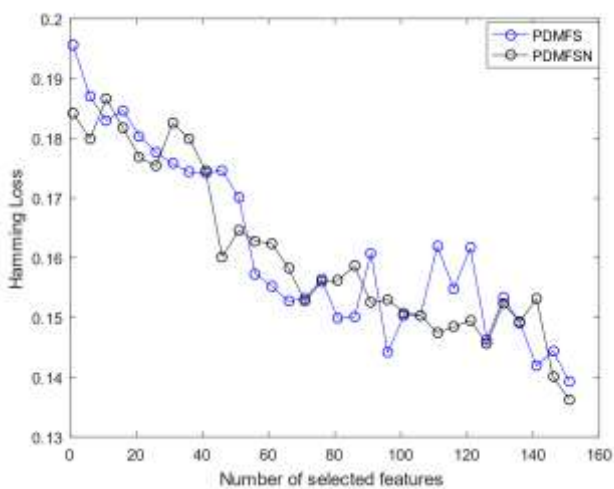
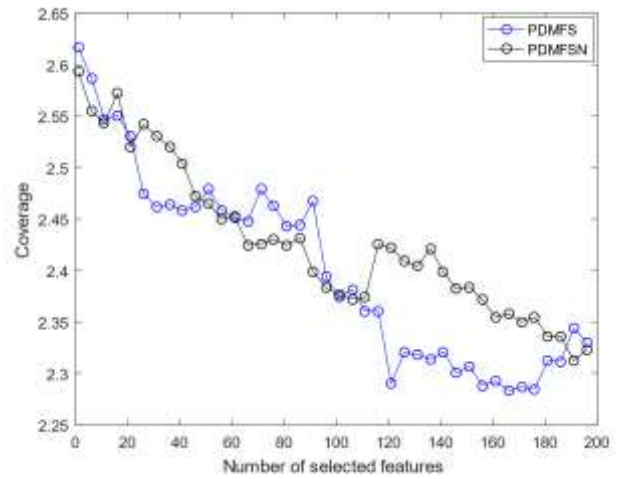
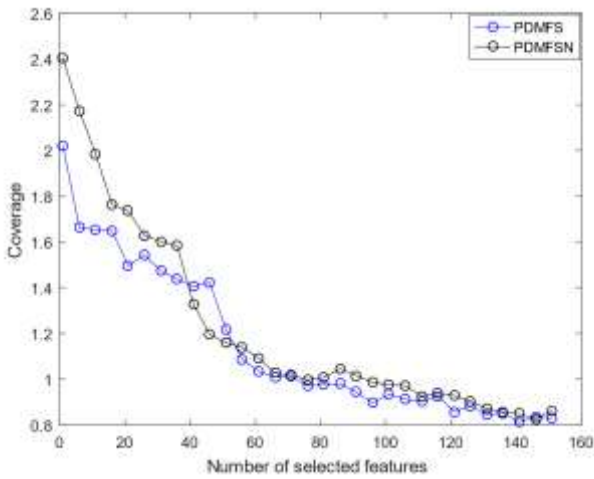
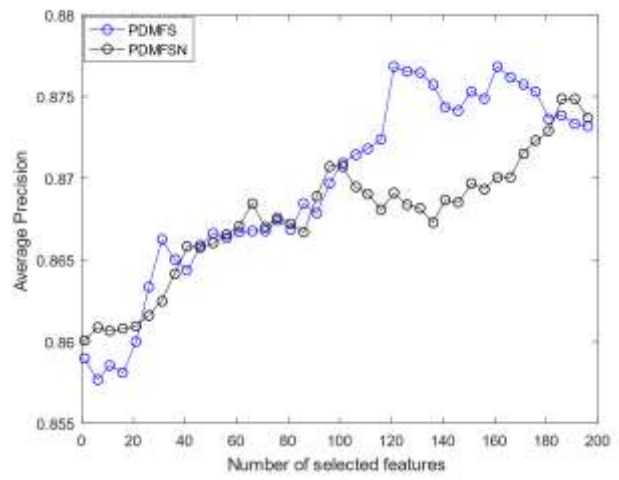
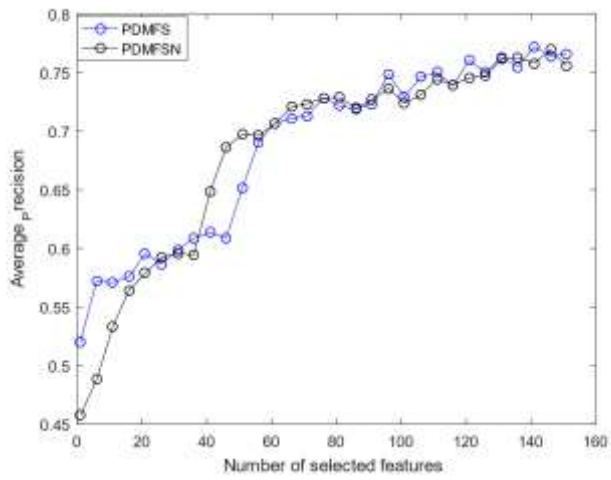
Dataset	MDDM _{spc}	MDDM _{proj}	PMU	MFNMI _{pes}	MFS-MCDM	MCMFS	PDMFS
Scene	0.2928	0.3383	0.3914	0.2823	0.4555	0.4964	0.4588
Business	0.0521	0.0547	0.0417	0.0729	0.0622	0.0894	0.0925
Computers	0.0803	0.0803	0.0767	0.0761	0.0547	0.0780	0.0878
Education	0.0598	0.0614	0.0376	0.0292	0.0084	0.0399	0.0646
Health	0.1506	0.1442	0.0969	0.0333	0.0752	0.1701	0.1545
Recreation	0.0590	0.0559	0.0909	0.1114	0.0427	0.1269	0.1301
Reference	0.0502	0.0535	0.0461	0.0579	0.0221	0.0573	0.0715
Science	0.0396	0.0375	0.0332	0.0301	0.0102	0.0404	0.0598
<i>Average</i>	4.1875	4.0625	5.1250	5.0000	6.0000	2.3750	1.2500

Table 7. Micro-F1 performance for seven feature selection methods (\uparrow).

Dataset	MDDM _{spc}	MDDM _{proj}	PMU	MFNMI _{pes}	MFS-MCDM	MCMFS	PDMFS
Scene	0.3318	0.3510	0.4243	0.3095	0.5017	0.5056	0.5030
Business	0.6768	0.6754	0.6721	0.6841	0.6746	0.6914	0.6951
Computers	0.3870	0.3870	0.3657	0.3942	0.3534	0.4087	0.4453
Education	0.1569	0.1439	0.0950	0.0596	0.0392	0.1108	0.1967
Health	0.3776	0.3477	0.3595	0.0726	0.3488	0.4938	0.4281
Recreation	0.0821	0.0835	0.1707	0.1908	0.0507	0.2468	0.2393
Reference	0.3443	0.3561	0.3126	0.3596	0.1900	0.3421	0.3890
Science	0.0904	0.0839	0.0931	0.0936	0.0274	0.1051	0.1376
<i>Average</i>	4.3125	4.6875	5.0000	4.2500	6.1250	2.2500	1.3750

4. For the ranking loss in Table 3, PDMFS obtains the best results on six datasets. Compared with the other four datasets, the difference between the score obtained by PDMFS and the best score is very small. For example, on the Health dataset, MCMFS obtains the best result, which is only 0.0004 better than the result obtained by PDMFS.
5. According to the results in Table 3 and Table 7, PDMFS obtains the best scores on six datasets in Macro-F1 score. In terms of the Micro-F1 score, PDMFS obtains the best scores on five datasets. For the other tree datasets, the maximum difference between the results obtained by PDMFS and other algorithms is 0.0657, and the minimum difference is only 0.0026.

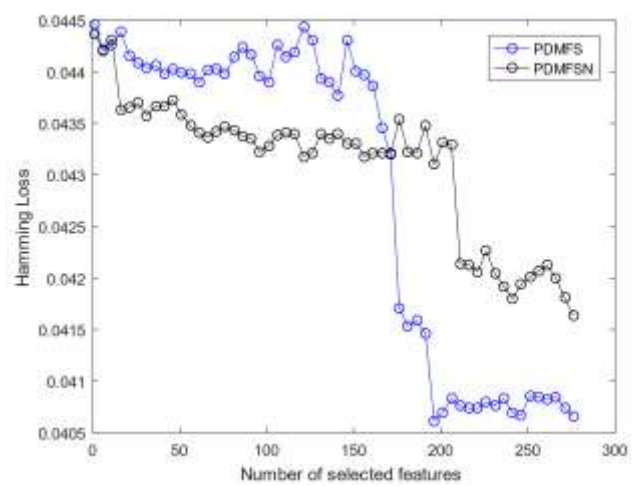
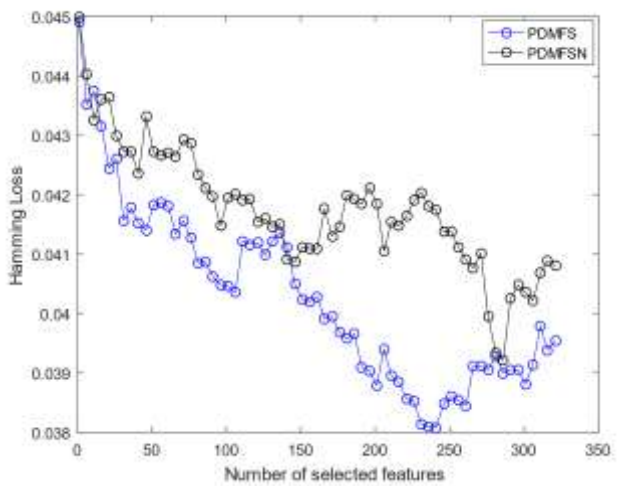
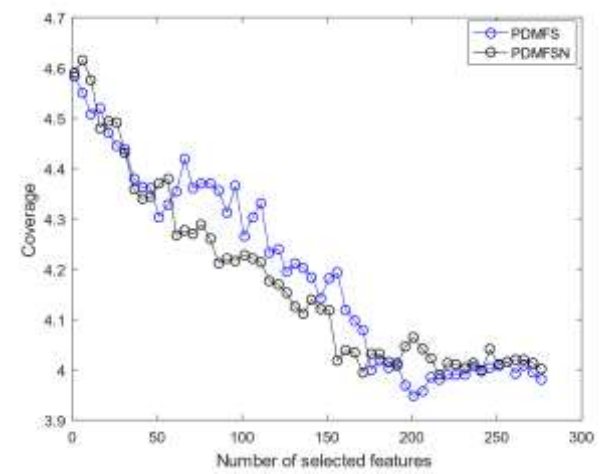
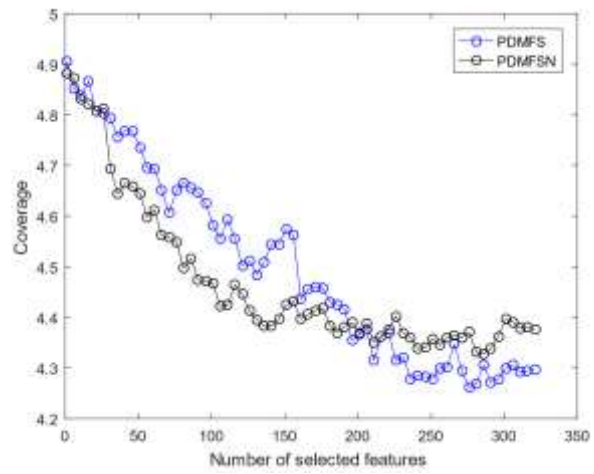
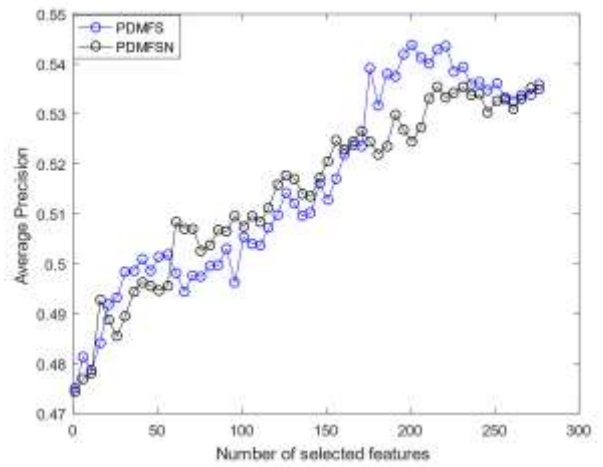
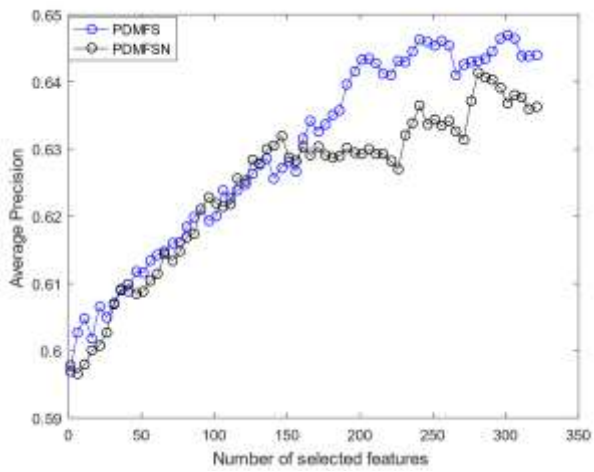
6. Overall, PDMFS performs best among other algorithms in terms of the average ranking results on the six indicators. It is clear that feature selection methods that consider multiple correlations simultaneously are superior to those with single correlation.



(a) Scene

(b) Business

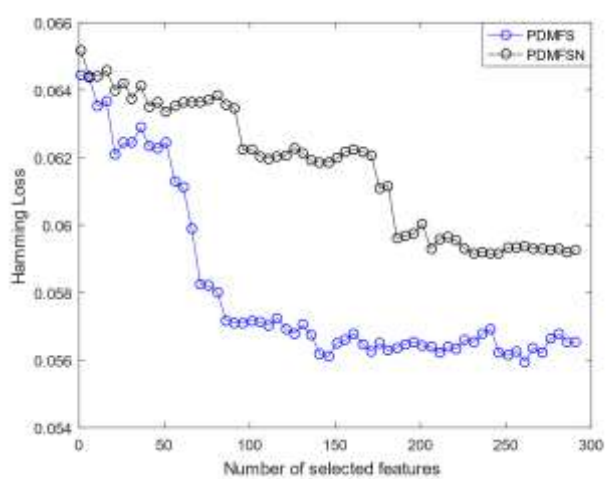
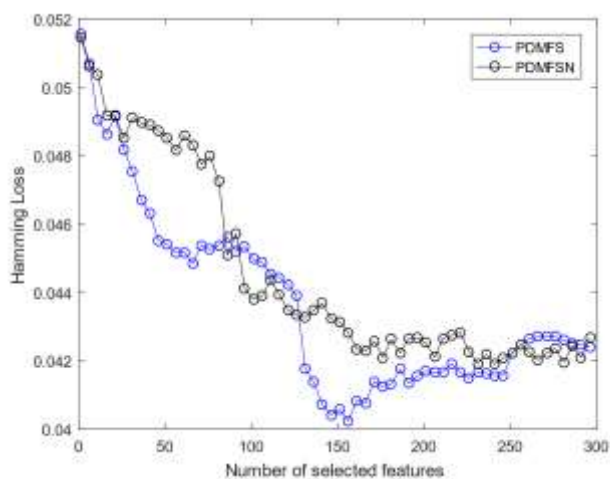
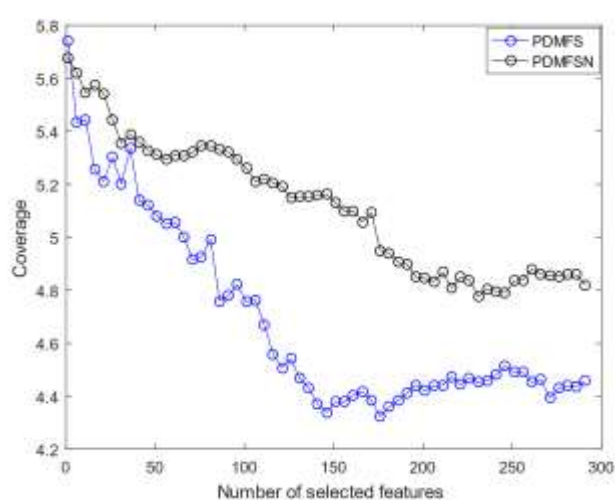
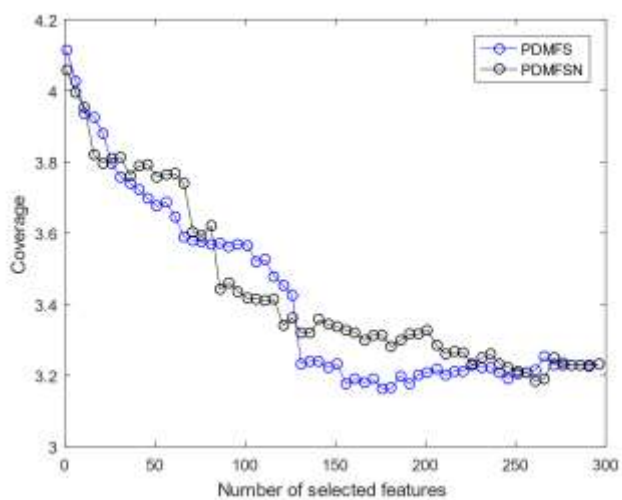
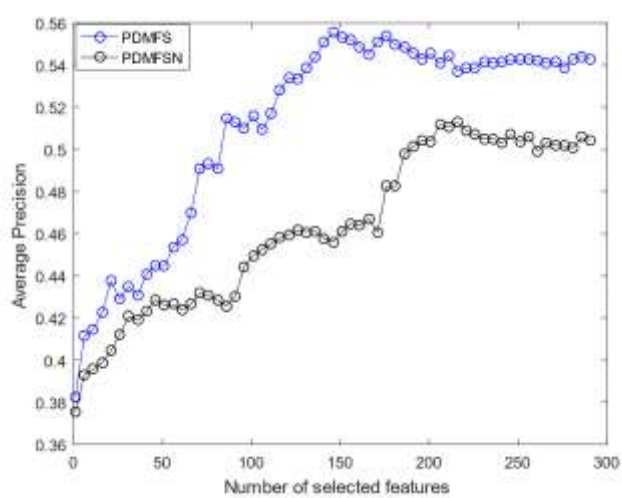
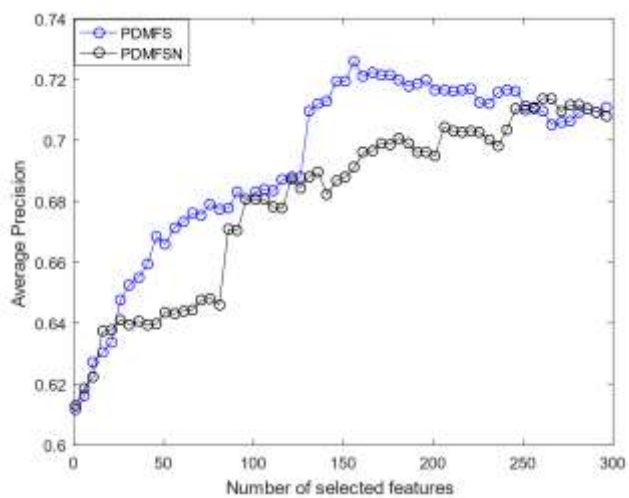
Figure 3. Experimental results on Scene and Business datasets.



(a) Computers

(b) Education

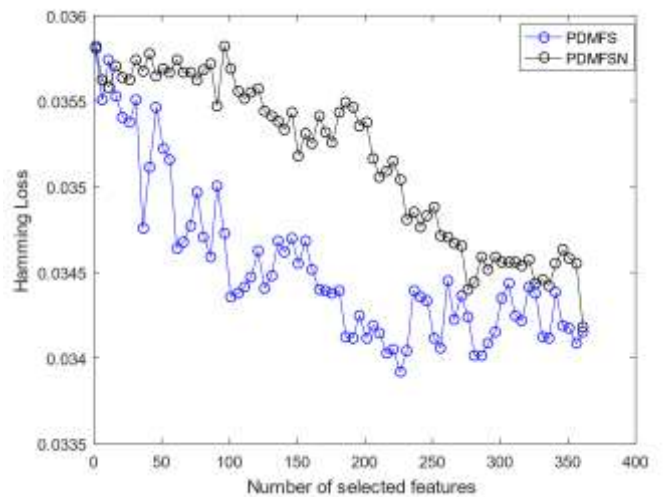
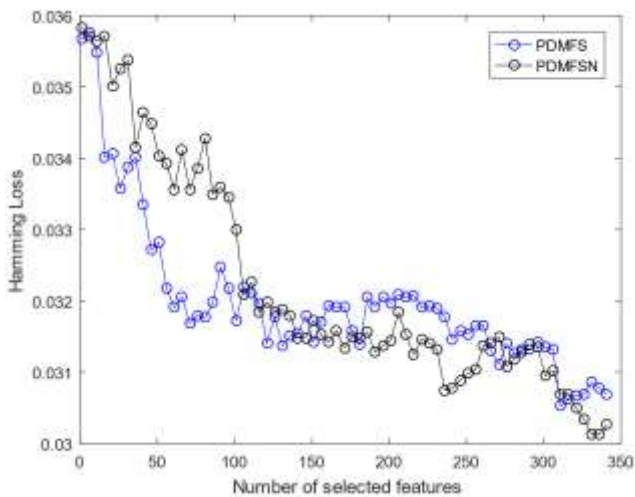
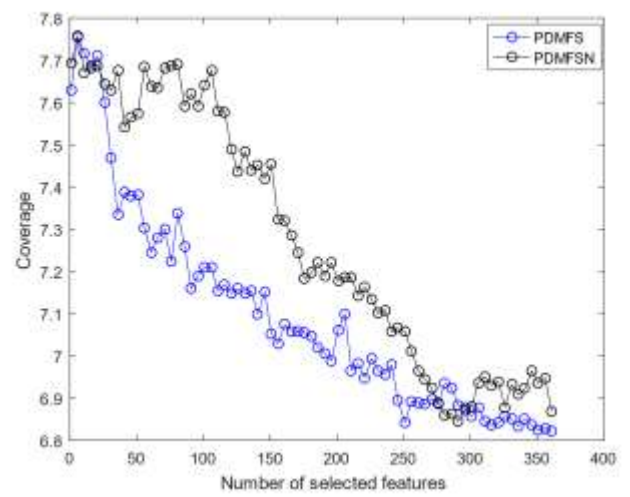
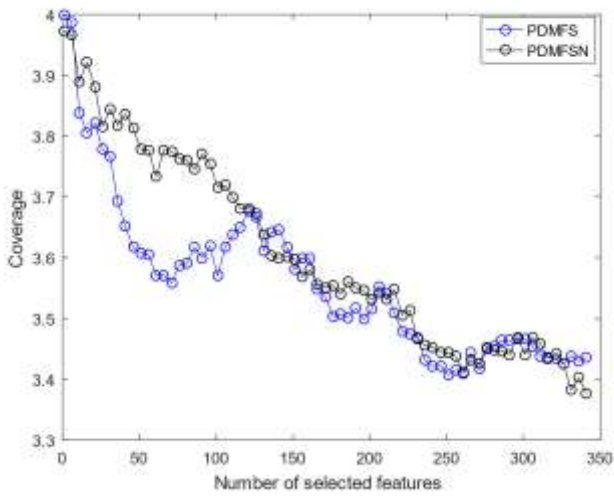
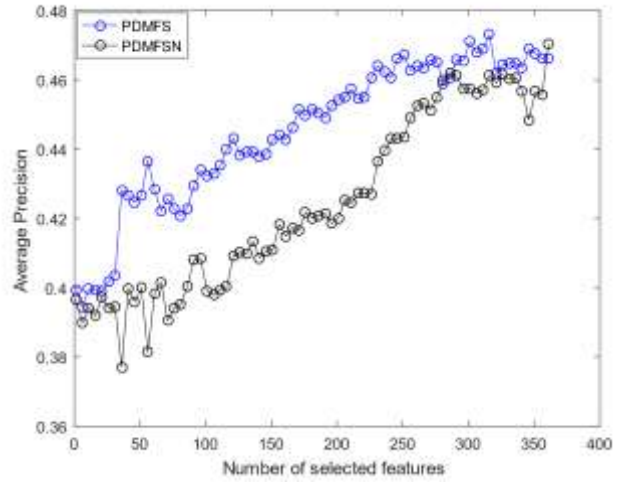
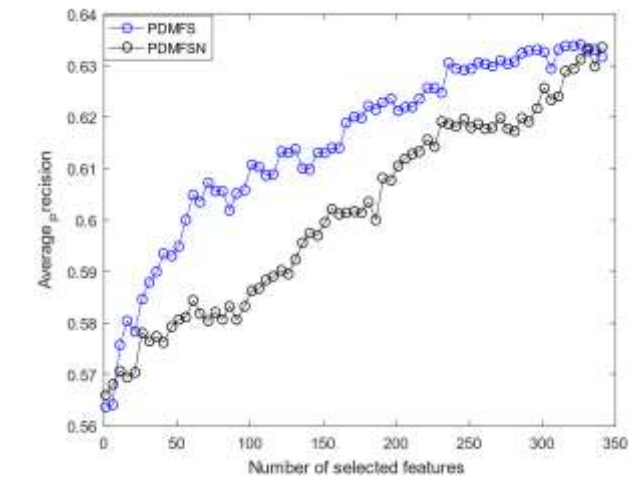
Figure 4. Part results on Computers and Education datasets.



(a) Health

(b) Recreation

Figure 5. Part results on Health and Recreation datasets.



(a) Reference

(b) Science

Figure 6. Part results on Reference and Science datasets.

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 3-6 and the interval for the number of features is five.

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
<i>Metric</i>	AP (\uparrow)		CV (\downarrow)		HL (\downarrow)	

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 **Error! Reference source not found.** (a) Scene and Figure Figure 6 (a) Reference, 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.

Most of the comparisons in **Error! Reference source not found.**-Figure 6 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 in Figure Figure 4 (b) Education. 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.

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:

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

where $K = 7, N = 8, q_\alpha = 2.9480, CD = 3.1842$. Figure 7 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 36 kinds of experimental comparison results (6 comparison algorithms, 6 evaluation indicators). The following observations are obtained from the results shown in Figure 7. We can conclude that PDMFS ranks No.1 among all methods, and PDMFS is significantly different from the other algorithms in 63.9% of cases. The results show that PDMFS has a high competitive performance compared to the comparative algorithms.

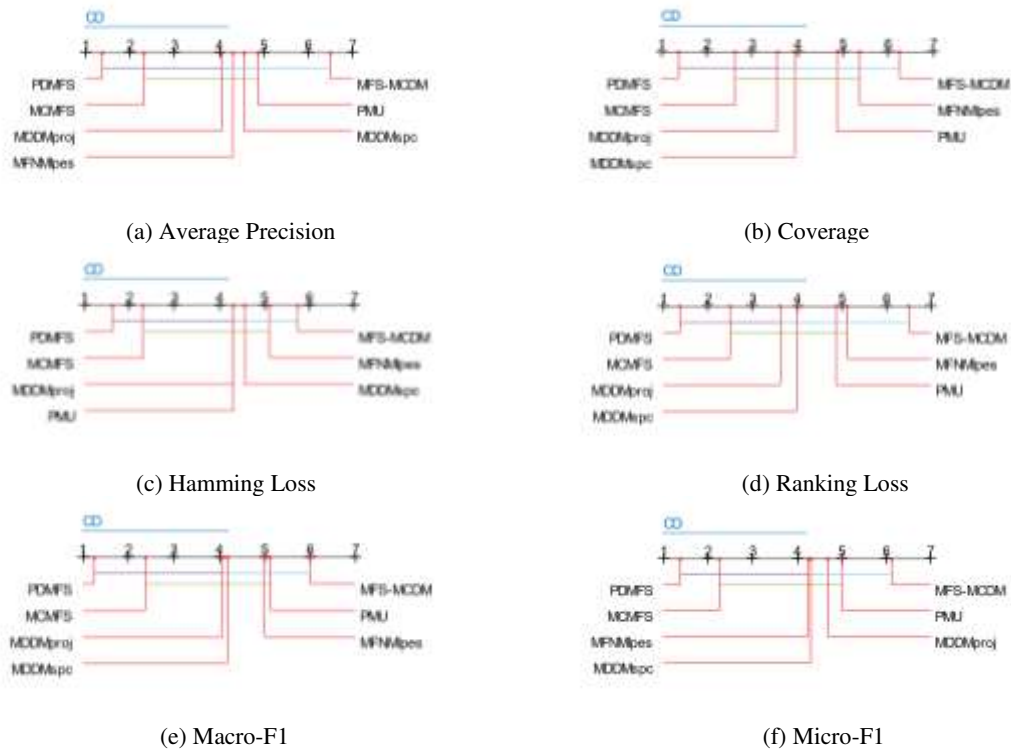


Figure 7. The performance comparison of algorithms

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

-
9. Hu, L.; Li, Y.; Gao, W.; Zhang, P.; Hu, J. Multi-label feature selection with shared common mode. *Pattern Recognition*. 2020, 104, 107344.
 10. Jiang, L.; Yu, G.; Guo, M.; Wang, J. Feature selection with missing labels based on label compression and local feature correlation. *Neurocomputing*. 2020, 395, 95-106.
 11. Zhang, P.; Gao, W.; Hu, J.; Li, Y. Multi-label feature selection based on high-order label correlation assumption. *Entropy*. 2020, 22, 797.
 12. Wang, Y.; Zheng, W.; Cheng, Y.; Zhao, D. Joint label completion and label-specific features for multi-label learning algorithm. *Soft Computing*. 2020, 24, 6553-6569.
 13. Cheng, Y.; Zhang, C.; Pang, S. Multi-label space reshape for semantic-rich label-specific features learning. *Int. J. Mach. Learn. & Cyber.* 2021, (*Published*).
 14. Wang, Y.; Zheng, W.; Cheng Y.; Zhao, D. Two-level label recovery-based label embedding for multi-label classification with missing labels. *Applied Soft Computing*, 2020, 99,106868.
 15. Huang, J.; Li, G.; Huang, Q.; Wu, X. Learning label specific features for multi-label classification. Proceedings of the IEEE International Conference on Data Mining. IEEE, Atlantic City, New Jersey, USA, November 14-17, 2015. pp: 181-190.
 16. Zhang, L.; Cheng, T.; Wang, Y.; Pei, G. Feature-label dual-mapping for missing label-specific features learning. *Soft Computing*. 2021, 25, 9307-9323.
 17. Cui, X.; Zou, C.; Wang, Z. Remote sensing image recognition based on dual-channel deep learning network. *Multimedia Tools and Applications*. 2021, 80, 27683-27699.
 18. Li, H.; Zheng, Y.; Ren, P. Dual-channel attention model for text sentiment analysis. *International Journal of Performability Engineering*. 2019, 15 ,834-841.
 19. Zhou, F.; Ma, Y.; Wang, B.; Lin, G. Dual-channel convolutional neural network for power edge image recognition. *Journal of Cloud Computing*. 2021, 10, 10-18.
 20. Wang, X., Liu, Y.; Du, Z.; Zhu, M.; Kaushik, A.C.; Jiang, X. Wei, D. Prediction of protein solubility based on sequence feature fusion and DDCNN. *Interdisciplinary sciences: computational life sciences*. 2021, 13, 703-716.
 21. Xu, Y.; Lu, L.; Xu, Z.; He, J.; Zhou, J.; Zhang, C. Dual-channel CNN for efficient abnormal behavior identification through crowd feature engineering. *Machine Vision and Applications*. 2019, 30, 945-958.
 22. Lee, J.; Kim, D.W. Memetic feature selection algorithm for multi-label classification. *Information Sciences*. 2015, 293, 80-96.
 23. Lee, J.; Kim, D.W. Mutual information based multi-label feature selection using interaction information. *Expert Systems with Applications*. 2015, 42, 2013-2025.
 24. Lin, Y.; Hu, X.; Wu, X. Quality of information-based source assessment and selection. *Neurocomputing*. 2014, 133, 95-102.
 25. Estrela, G.; Gubitoso, M.D.; Ferreira, C.E.; Barrera, J.; Reis, M.S. An Efficient, Parallelized Algorithm for Optimal Conditional Entropy-Based Feature Selection. *Entropy*, 2020, 22, 492.
 26. Lee, J.; Kim, D.W. Feature selection for multi-label classification using multivariate mutual information. *Pattern Recognition Letters*. 2013, 34, 349-357.
 27. Zhang, Y.; Zhou, Z.H. Multi-label dimensionality reduction via dependence maximization. Proceedings of the 23rd AAAI Conference on Artificial Intelligence. AAAI Press, Chicago, Illinois, USA, July 13-17, 2008. pp: 1503-1505
 28. Amin, H.; Mohammad, B.D.; Hossein, N. MFS-MCDM: Multi-label feature selection using multi-criteria decision making. *Knowledge-Based Systems*. 2020, 206, 106365.
 29. Lin, Y.; Hu, Q.; Liu, J. Multi-label feature selection based on max-dependency and min-redundancy. *Neuro Computing*. 2015, 168, 92-103.
 30. Yang, M.; Wang, F. A classifier ensemble algorithm based on local random subspace. *Pattern Recognition and Artificial Intelligence*. 2012, 25, 595-603.
 31. Liu, J.; Lin, M.; Wang, C.; Lin, Y. Multi-label feature selection algorithm based on local subspace. *Pattern Recognition and Artificial Intelligence*. 2016, 29, 240-251.
 32. Lin, Y.; Hu, Q.; Liu, J.; Chen, J.; Duan, J. Multi-label feature selection based on neighborhood mutual information. *Applied Soft Computing*, 2016, 38,244-256.
 33. Schapire, R.E.; Singer, Y. BoosTexter: A Boosting-based System for Text Categorization. *Mach. Learn.* 2000, 39, 135-168.
 34. Zhang, M.L.; Zhou, Z.H. ML-KNN: A lazy learning approach to multi-label learning[J]. *Pattern Recognition*.2007, 40, 2038-2048.
 35. Janez, D.; Dale, S. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*. 2006, 7, 1-30.