Mathematical Biosciences and Engineering

MBE, 20(8): 14201–14221.
DOI: 10.3934/mbe.2023635
Received: 27 April 2023
Revised: 10 June 2023
Accepted: 20 June 2023
Published: 26 June 2023

http://www.aimspress.com/journal/MBE

Research article

Multi-label feature selection based on HSIC and sparrow search algorithm

Tinghua Wang*, Huiying Zhou and Hanming Liu

School of Mathematics and Computer Science, Gannan Normal University, Ganzhou 341000, China

* **Correspondence**: Email: wthpku@163.com.

Abstract: Feature selection has always been an important topic in machine learning and data mining. In multi-label learning tasks, each sample in the dataset is associated with multiple labels, and labels are usually related to each other. At the same time, multi-label learning has the problem of "curse of dimensionality". Feature selection therefore becomes a difficult task. To solve this problem, this paper proposes a multi-label feature selection method based on the Hilbert-Schmidt independence criterion (HSIC) and sparrow search algorithm (SSA). It uses SSA for feature search and HSIC as feature selection criterion to describe the dependence between features and all labels, so as to select the optimal feature subset. Experimental results demonstrate the effectiveness of the proposed method.

Keyword: feature selection; multi-label classification; sparrow search algorithm; Hilbert-Schmidt independence criterion (HSIC); data mining

1. Introduction

Feature selection aims at determining a subset of available features which is most discriminative and informative for data analysis [1]. In practical applications, more features may result in higher data collection costs, greater difficulty in model interpretation, higher computational costs for the predictor, and sometimes lower generalization capabilities [2]. Therefore, it is important to perform feature selection before actual learning.

Label data usually has thousands or even tens of thousands of features, especially images and texts; and each instance is associated with multiple labels at the same time. For example, a gene in

bioinformatics may be related to multiple functions. Each document in information retrieval may cover multiple topics. In image processing, images can be annotated with different scenes. For a given learning task, many features are redundant and irrelevant. High-dimensional data may bring many shortcomings to the learning algorithm, such as the large amount of calculation, overfitting and poor performance. To solve this problem, researchers have proposed multi-label feature selection algorithms to reduce the dimension of multi-label data, improve the accuracy of classification learning, and generate more compact and generalized classification models. Therefore, multi-label feature selection is an important research topic in pattern recognition, machine learning and other fields. For example, in the biomedicine field, multi-label feature selection is widely used in case data analysis to extract various information contained in cancer data and improve the cure rate of cancer. In the financial field, multi-label feature selection helps financial companies better recommend funds to users.

Hilbert-Schmidt independence criterion (HSIC) [3] is a measure of the strength of dependence between two variables and is the most common kernel statistical independence criterion, including biased HSIC and unbiased HSIC versions. At present, due to the effectiveness and low computational complexity of the criterion, it is widely used in various machine learning problems, such as clustering [4], dimensionality reduction [5], feature selection [6], independent component analysis (ICA) [7] and canonical correlation analysis (CCA) [8]. When HSIC is applied to the feature selection problem, it can describe the dependency between the selected features and all labels.

The swarm intelligence algorithm is a random search algorithm inspired by social behavior patterns, evolution mechanisms and physical phenomena of biological groups in nature [9]. It includes the ant colony optimization (ACO) [10,11], particle swarm algorithm (PSO) [12–14], grey wolf optimization (GWO) [15,16], sparrow search algorithm (SSA) [17], etc. Among them, SSA is a new swarm intelligence optimization algorithm proposed by Xue et al. in 2020. Compared with other swarm intelligence optimization algorithms, it has the characteristics of high search accuracy, fast convergence speed, good stability and strong robustness.

At present, many researchers have studied multi-label feature selection. Sun et al. [18] proposed an improved ReliefF multi-label feature selection algorithm based on global sample correlation, but it only relies on the correlation between features and labels to select feature subsets, ignoring the dependence between labels. Mutual information (MI) is a dependency measure of variables, and it can be used to assess the correlation of variables [19]. González-López et al. [20] proposed two multi-label feature selection methods based on minimum redundancy and maximum relevance. Some researchers extended mutual information to fuzzy mutual information for feature selection. For example, Xiong et al. [21] proposed a feature selection algorithm based on label distribution and fuzzy mutual information. Some researchers also extended mutual information to conditional mutual information (CMI). Sha et al. [22] proposed a new filtering feature selection method based on CMI, and the experimental results show that this method has advantages in label prediction. When HSIC is applied to the feature selection problem, it can describe the dependency between the selected features and all labels. Liu et al. [23] proposed a multi-label feature selection method based on the unbiased HSIC and control genetic algorithm, but it does not consider the dependency between labels. Li et al. [24,25] proposed two multi-label feature selection methods with Pareto optimality for continuous data, but neither of them is suitable for the case of few features and many labels, and they do not analyze the correlation between labels.

Feature selection can be implemented by swarm intelligence algorithms. Paniri et al. [10]

proposed a multi-label feature selection algorithm based on ACO (MLACO). By introducing two unsupervised and supervised heuristic functions, the features with low redundancy and high relevance to class labels were found. Experimental results show that the method has better classification performance. Paniri et al. [11] proposed a multi-label feature selection algorithm combining ACO and time difference reinforcement learning. The algorithm can achieve better classification performance by using the heuristic function of reinforcement learning ACO. Zhang et al. [12] proposed a wrapper multi-label multi-objective feature selection algorithm based on the PSO algorithm. This method used a probability-based coding strategy to represent each particle, which makes the problem suitable for PSO. Different from the off-line multi-label feature selection methods based on PSO, Paul et al. [13] proposed a multi-objective multi-label online feature selection method based on PSO. However, the method does not consider the dependence between labels, and if a large number of significant features appear before feature selection, it may make the algorithm fall into incalculable difficulties. Feature selection based on swarm intelligence algorithms has the disadvantages of large computation, being easy to fall into local optima, slow convergence speed and bad classification performance.

This paper proposes a multi-label feature selection method based on HSIC and SSA (MLSSA). This method searches features according to SSA, and uses HSIC as a feature selection criterion to describe the dependence between features and all labels, and then selects the optimal feature subset. The performance of the proposed method is evaluated by experiments on eight datasets. The results of different evaluation indicators show that the proposed method can improve the classification performance and is superior or competitive to other comparison methods. SSA has been used for single-label feature selection [26], but there is no report on the application of SSA to multi-label feature selection. Therefore, the proposed algorithm is the first to apply SSA to the field of multi-label feature selection, and for the first time to use HSIC in the fitness function to distinguish good and bad individuals in the sparrow population. The main contributions of this paper are outlined as follows:

• A multi-label feature selection method based on HSIC and SSA was proposed, which uses SSA for feature search and HSIC as feature selection criterion to describe the dependence between features and all labels.

• To the best of our knowledge, this is the first time that SSA is used for multi-label feature selection and the HSIC is used as the fitness function of SSA.

• Comprehensive experiments on real-world datasets verify the effectiveness of the proposed MLSSA method.

The rest of this article is summarized as follows. Section 2 describes the relevant knowledge, including HSIC and the sparrow search algorithm. Section 3 introduces the proposed multi-label feature selection method in detail. Section 4 discusses the experimental results. Section 5 draws conclusions.

2. Preliminaries

Suppose the number of samples is m, the number of labels is q, X is the d dimension instance space $\mathbb{R}^d$; Y is a set of labels with q possible class labels $Y = \{y_j \mid j = 1, \ldots, q\}$. The task of multi-label learning is to learn a function $h: X \to 2^Y$ from the multi-label training set

$D=\{(\boldsymbol{x}_i,y_i)|i=1,\ldots,m\}$. For any unknown instance E_i ($\boldsymbol{x}_i$ is a d dimensional feature vector, Y_i is a label set related to $\boldsymbol{x}_i$), the multi-label classifier $h(\bullet)$ predicts that $h(\boldsymbol{x}_i)\subseteq Y$ is the appropriate label set of $\boldsymbol{x}_i$.

2.1. HSIC

HSIC is an independence measure based on kernel functions. An independence criterion is obtained by calculating the empirical estimate of the Hilbert-Schmidt cross-covariance operator norm between variables in the reproducing kernel Hilbert space (RKHS). The empirical estimation of HSIC has been proved to have the advantages of fast convergence speed and simple calculation (computational complexity is $o(m^2)$) in theory. The greater the value, the stronger the correlation between X and Y is, and a value of 0 indicates that X and Y are independent of each other.

Let F be the RKHS of the X to R, where X and R are the metric space and a set of real numbers respectively. For a point $x,x^{'}\in X$, there is a corresponding element $\phi(x),\phi(x^{'})\in F$ (we call $\phi:X\to F$ as a feature map) with $k(x,x^{'})=\left\langle\phi(x),\phi(x^{'})\right\rangle_F$, where $k:X\times X\to R$ is the related reproducing kernel. Let G be the RKHS of function Y to R, where Y is the metric space with feature maps $\varphi:Y\to G$ and $l(y,y^{'})=\left\langle\varphi(y),\varphi(y^{'})\right\rangle_G$, where $y,y^{'}\in Y$.

Let $(x,y)\subseteq X\times Y$ be a random variable derived from the joint probability distribution P_{xy}, then the covariance matrix can be defined as:

$$C_{xy}=\mathrm{E}_{xy}(xy^{\mathrm{T}})-\mathrm{E}_x(x)\mathrm{E}_y(y^{\mathrm{T}}) \tag{1}$$

where E_{xy},E_x,E_y are the expected values of the probability distributions P_{xy}, P_x and P_y, respectively, and y^{T} is the transpose of y. The Frobenius norm can effectively generalize the degreeof linear correlation between x and y:

$$\|C_{xy}\|_{\mathrm{Frob}}=\|C_{xy}\|_{\mathrm{HS}}=\sqrt{\mathrm{tr}(C_{xy}C_{xy}^{\mathrm{T}})} \tag{2}$$

where $tr(\bullet)$ is the trace operator, which is 0 if and only if there is no linear correlation between x and y, so it can be used to detect a linear correlation between them. However, such statistics are fairly limited [27,28].

In order to address these limitations, the concept of the Frobenius norm is extended to HSIC: Given two feature mappings $\phi:X\to F$ and $\varphi:Y\to G$, the linear operator $C_{xy}:G\to F$ is the cross-covariance operator between ϕ and φ, such that:

$$C_{xy}=\mathrm{E}_{x,y}[[\phi(x)-\mathrm{E}_x[\phi(x)]]\otimes[\varphi(y)-\mathrm{E}_y[\varphi(y)]]] \tag{3}$$

where $\otimes$ is the tensor product. The square of the Hilbert-Schmidt norm of the cross-covariance operator can be defined as HSIC:

$$\begin{aligned}HSIC(\mathrm{F},\mathrm{G},\mathrm{P}_{xy})=\|C_{xy}\|_{\mathrm{HS}}^2&=\mathrm{E}_{xx'yy'}[k(x,x')l(y,y')]\\&-2\mathrm{E}_{xy}[\mathrm{E}_{x'}[k(x,x')]\mathrm{E}_{y'}[l(y,y')]]\\&+\mathrm{E}_{xx'}[k(x,x')]\mathrm{E}_{yy'}[l(y,y')]\end{aligned} \tag{4}$$

where $E_{xx'yy'}$ is the expected value of $(x,y)\sim P_{xy}$ and $(x',y')\sim P_{xy}$. It shows that the Hilbert-Schmidt norm exists when the kernels k and l are bounded. If both feature maps are linear, then HSIC is the same as the second power of the Frobenius norm [29,30].

Given a set $D=\{(x_i,y_i)\}_{i=1}^{m}$ from P_{xy} and the selected kernel k and l, we can form two kernel matrices $\mathbf{K}$, $\mathbf{L}\in \mathrm{R}^{m\times m}$, where $\mathbf{K}_{ij}=k(x_i,x_j)$, $\mathbf{L}_{ij}=l(y_i,y_j)$, and it takes the following form:

$$HSIC(\mathrm{F},\mathrm{G},D)=\frac{1}{(m-1)^2}\mathrm{tr}(\mathbf{KHLH}) \tag{5}$$

where $\mathbf{H}=\mathbf{I}_m-\boldsymbol{e}_m\boldsymbol{e}_m^{\mathrm{T}}/m\in \mathrm{R}^{m\times m}$ is the central matrix, $\mathbf{I}_m\in \mathrm{R}^{m\times m}$ and $\boldsymbol{e}_m\in \mathrm{R}^{m}$ are the unit matrix and vector of 1, respectively.

2.2. Sparrow search algorithm

The SSA algorithm is an optimal search strategy designed for the search and anti-predation characteristics of sparrow. Its basic principle is that the search process can be summarized as a discoverer-entrant model, and the reconnaissance and early warning mechanism is incorporated. It is found that individuals are highly adaptable and have a wide range of search capabilities, which can guide group search and foraging. In order to better adapt to the environment, the attender will follow the discoverers for foraging. In addition, in order to increase their ability to hunt, some individuals monitor the finders in order to compete for food or search for food around them [31].

Suppose that there are N sparrows in d dimensional search space, and the position of a population of N sparrows in a d-dimensional space is $P_i=[p_1,\ldots,p_N]$, $p_i=[p_{i,1},\ldots,p_{i,d}]$, where $i=1,2,\ldots N$, $p_{i,d}$ represents the position of the ith sparrow in the d dimension search space. The adaptation values of the sparrow are $F_p=[f(p_1),\ldots f(p_i),\ldots f(p_N)]^T$ and $f(p_i)=[f(p_{i,1}),f(p_{i,2}),\ldots f(p_{i,d})]$, where $i=1,2,\ldots N$. Each value in F_p represents the fitness value of the individual. In SSA, producers with better fitness values preferentially obtain food during the search process.

Discoverers generally account for 10 to 20% of the population, and the location update formula is as follows:

$$p_{i,d}^{t+1}=\begin{cases} p_{i,d}^{t}\cdot\exp\left(\dfrac{-i}{\alpha\cdot T}\right) & ,R_2<ST \\ p_{i,d}^{t}+Q\cdot L & ,R_2\geq ST \end{cases} \tag{6}$$

where t is the current iteration number, T is the maximum iteration number, and $p_{i,d}^{t}$ is the dth dimensional value of the ith sparrow at the tth iteration. α is a uniform random number between $(0,1]$. Q is a random number obeying standard normal distribution. L represents a matrix of size $1\times d$ and elements 1. $R_2\in[0,1]$ and $ST\in[0.5,1]$ represent the warning value and the safety value, respectively. When $R_2<ST$, the population does not find the existence of predators or other dangers, the search environment is safe, and discoverers can be widely searched to guide the population to obtain a higher fitness. When $R_2\geq ST$, sparrows detect predators and immediately

release danger signals. The population immediately performs anti-predation behavior, adjusts the search strategy, and quickly moves closer to the safe area.

In addition to the discoverer, the remaining sparrows are the entrants and the location is updated as follows:

$$p_{i,d}^{t+1} = \begin{cases} Q \cdot \exp(\frac{p_{worst}^{t} - p_{i,d}^{t}}{i^2}) & ,i > \frac{n}{2} \\ p_b^{t+1} + \left| p_{i,d}^{t} - p_b^{t+1} \right| \cdot A^{+} \cdot L & ,otherwise \end{cases} \tag{7}$$

where p_{worst}^{t} represents the worst position of the sparrow at the tth iteration of the population. p_b^{t+1} represents the optimal position of the sparrow in the $(t+1)$th iteration. A represents that each element value of a $1 \times d$ matrix is randomly assigned 1 or -1, $A^{+} = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}$. When $i > \frac{n}{2}$, it indicates that the ith participant does not receive any food, and then it is hungry and has low fitness, so in order to gain more energy they must fly to other places to search food. Otherwise, the ith participant will randomly find a position near the current optimal position p_b for foraging.

Reconnaissance warning sparrows generally account for 10 to 20% of the population, and the location update is as follows:

$$p_{i,d}^{t+1} = \begin{cases} p_{best}^{t} + \beta(p_{i,d}^{t} - p_{best}^{t}) & ,f_i \neq f_g \\ p_{i,d}^{t} + K\left(\frac{\left| p_{i,d}^{t} - p_{worst}^{t} \right|}{(f_i - f_w) + e}\right) & ,f_i = f_g \end{cases} \tag{8}$$

where p_{best} is the current optimal position of the population. β is a random number, and follows the normal distribution of mean 0 and variance 1. It is also a step size control parameter. $K \in [-1,1]$ represents both the direction in which the sparrow moves and the step size control coefficient. To avoid having a denominator of 0, e is represented as a minimal constant. f_i represents the fitness value of the current individual, f_g and f_w represent the best and worst fitness values of the current population, respectively. When $f_i \neq f_g$, it indicates that the individual is located in the edge population and is vulnerable to natural enemies. When $f_i = f_g$, it represents an individual located in the center of the population, which senses the threat of natural enemies and approaches other individuals in time to avoid being attacked by natural enemies.

3. Proposed method

In this section, HSIC is used to describe the dependence between features and labels and SSA is used for feature search. A multi-label feature selection algorithm MLSSA based on HSIC and SSA is proposed. The process is shown in Table 1. First, the algorithm initializes the sparrow population parameters. Second, the HSIC value of each feature is calculated and stored in the fitness value, and the fitness value is sorted to find the optimal position and the worst position. Then, the current sparrow position is updated and obtained, and it is compared with the previous position. If it is better than the previous position, the fitness value and the optimal position are updated. Otherwise, the

sparrow position is continuously updated until the maximum number of iterations is reached. Finally, the recorded positions and fitness values are sorted in descending order, and the top n features are extracted as the optimal feature subset.

Table 1. MLSSA algorithm.

Algorithm 1. The pseudo-codes for MLSSA	
Input	X : feature data matrix; Y : label data matrix; n : the number of selected features; G : the maximal number of generations; N : Sparrow population size; Kernel function types and parameters of label data; Proportion of sparrow population discoverers, producers, warnings, warning values R_2 ;
1	Initialize sparrow population, set population size, evolution times, warning value, discoverer, producer ratio;
2	Calculate the kernel matrix of label data;
3	Calculate the HSIC value for each feature and store it in the fitness value;
4	While （t < G）
5	Rank the fitness values to find the current optimal fitness value and the worst fitness value, and their corresponding position;
6	$R_2 = rand(1)$;
7	Update sparrow position according to Eqs (6)–(8);
8	Get the current position and compare it with the previous optimal position. If it is better than the previous optimal position, update and record;
9	$t = t + 1$;
10	end
11	The position and fitness value of the record are sorted in descending order, and the first n features are extracted;
12	These n features are the best subset of features;
Output	Optimal feature subset.

The specific flow chart is shown in Figure 1. For multi-label data, we first preprocess it, and then randomly select 300 data and divide them into training set and test set. We calculate the HSIC values between each feature and all labels and then put them into the fitness value, that is, we replace the fitness function with HSIC. We sort them to find the optimal fitness value and the optimal position. We then update and obtain the position and fitness value of the current sparrow. If they are better, we update the optimal position and fitness value until the maximum number of iterations is reached.

4. Experiments

4.1. Datasets

The datasets used in the experiment can be downloaded from the open source project mulan (http://mulan.sourceforge.net/datasets-mlc.html) or Multi-Label Classification Dataset Repository (https://www.uco.es/kdis/mllresources). These datasets are widely used in multi-label learning, as shown in Table 2. “Name” represents the name of the dataset, “Domain” represents the domain to which the dataset belongs, “Instances” represents the total number of samples of the dataset,

"Features" represents the total numbers of features of the dataset, "Labels" represents the total number of labels of the dataset and "Cardinality" represents the average category to which the samples of the dataset belong. These datasets cover different application fields. For example, the corel5k dataset contains 5000 corel images, each containing multiple segments such as cats, forests, grasslands and tigers, which are used for image classification scenarios. The genbase dataset is used for protein function classification, which belongs to biological classification scenarios. The rest of the datasets are widely used for text classification.

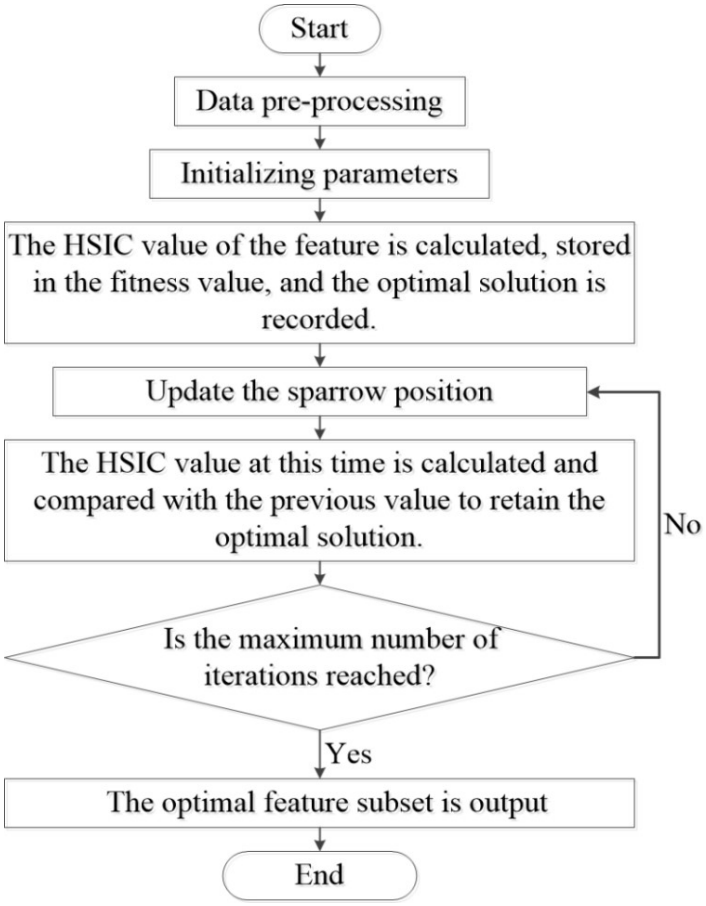

Figure 1. Algorithm flow chart.

Tabel 2. Datasets for multi-label learning.

Name	Domain	Instances	Features	Labels	Cardinality
corel5k	images	5000	499	374	3.522
tmc2007-500	text	28596	500	22	2.22
languagelog	text	1460	1004	75	1.180
medical	text	978	1449	45	1.245
enron	text	1702	1001	53	3.378
chess	text	1675	812	227	2.411
genbase	biology	662	1186	27	1.252
delicious	text (web)	16110	500	983	19.020

4.2. Evaluating indicators

The evaluation indicators including accuracy, precision, and recall in traditional single-label classification problems are not suitable for multi-label learning problems. The evaluation of multi-label learning problems is much more complicated than that of single-label learning. Literature [32] defines five commonly used evaluation indicators in multi-label learning, and the specific formula can be seen in the original text. The introduction is as follows:

1) Hamming loss

$$hloss(h) = \frac{1}{p}\sum_{i=1}^{p}\left|h(\boldsymbol{x}_i)\Delta Y_i\right| \tag{9}$$

where Δ represents the symmetry difference between two sets, and $\left|\bullet\right|$ returns the cardinality of the set $\bullet$. Hamming loss evaluates the percentage of misclassified instant-label pairs, that is, missing a relevant label or predicting an irrelevant label.

2) One-error

$$\text{One-error}(f) = \frac{1}{p}\sum_{i=1}^{p}[[\arg\max_{y\in L} f(\boldsymbol{x}_i, y)] \notin Y_i] \tag{10}$$

where the real-valued function $f: X \times L \to \mathbb{R}$, $f(\boldsymbol{x}, y)$ returns the confidence of the correct label of $\boldsymbol{x}$, and the one-error calculates the proportion of examples where the top-ranked label is not in the relevant label set.

3) Coverage

$$\text{coverage}(f) = \frac{1}{p}\sum_{i=1}^{p}\max_{y\in Y_i} rank_f(\boldsymbol{x}_i, y) - 1 \tag{11}$$

where $rank_f(\boldsymbol{x}, y)$ returns the order of y in L in the descending order of $f(\boldsymbol{x}, \bullet)$. The coverage evaluation takes on average how many steps to move the sorted label list down to cover all relevant labels of the example.

4) Ranking loss

$$\text{rloss}(f) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\left|Y_i\right|\left|\overline{Y}_i\right|}\left\{(y', y'')\middle| f(x_i, y') \le f(x_i, y'')\right\}, (y', y'') \in Y_i \times \overline{Y}_i \tag{12}$$

where $y' \in Y_i$ is the related label of $\boldsymbol{x}_i$, $y'' \notin Y_i$ is the unrelated label of $\boldsymbol{x}_i$ and $\overline{Y}$ is the

complementary set of Y. The ranking loss evalutes the proportion of reverse ranked label pairs, that is, the ranking of unrelated labels is higher than that of related labels.

5) Average precision

$$avgprec(f)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|}\sum_{y\in Y_i}\frac{\left|\left\{y'\middle|rank_f(\boldsymbol{x}_i,y')\le rank_f(\boldsymbol{x}_i,y),y'\in Y_i\right\}\right|}{rank_f(\boldsymbol{x}_i,y)} \tag{13}$$

Average precision evaluates the average score of related labels that are ranked above a particular label $y\in Y_i$.

For the above metrics (except average precision), the smaller the metric, the better the system performance is. The optimal value of coverage is $\frac{1}{p}\sum_{i=1}^{p}|Y_i|-1$, and the optimal value of one-error and ranking loss is 0. For the measure of average precision, the larger the measure, the better the system performs, and the optimal value is 1.

4.3. Comparison algorithms

The proposed multi-label feature selection algorithm is tested on the selected eight datasets, and compared with the following six algorithms: GRRO (multi-label feature selection method via global relevance and redundancy optimization) [33], GRRO-LS (GRRO with label-specific features) [33], FIMF (fast multi-label feature selection based on information-theoretic feature ranking) [34], PPT-MI (pruned problem transformation with mutual information) [35], PPT-CHI (pruned problem transformation with χ^2 test) [36] and Ant-TD (ant colony optimization plus temporal difference reinforcement learning for multi-label feature selection) [11].

GRRO: It is a multi-label feature selection method based on information theory. This method is a feature evaluation considering feature relevance, feature redundancy and label relevance, and the optimal solution can be obtained by processing the relevance and redundancy information once.

GRRO-LS: It is an extension of the algorithm GRRO. Considering that different labels have their inherent distinguishing features, the features selected by this method are label-specific.

FIMF: It is a fast multi-label feature selection method. It obtains a scoring function based on information theory to evaluate the importance of each feature, and then analyzes the results from the perspective of computational cost.

PPT-MI: It is a feature selection algorithm based on mutual information. The idea is to first use the PPT (pruned problem transformation) to transform the problem, and then use the greedy search algorithm based on MI to select the most relevant features.

PPT-CHI: It is the result obtained by Trochidis et al. using the method in [37], that is, the PPT method is used to transform the problem, and then the χ^2 statistic is used to rank the features.

ANT-TD: It is a multi-label feature selection method based on heuristic learning. This method uses the temporal difference reinforcement learning algorithm to learn heuristic function from experience, and combines the ant colony algorithm with heuristic learning.

4.4. Results and discussion

In this paper, multi-label k-nearest neighbor method (ML-kNN) is used as a classifier to calculate the above multi-label evaluation indicators, and the parameter k is set to 10. In this experiment, 300 data samples in the datasets are selected for experiments, of which 60% of the dataset samples are used for the training set, and the remaining 40% of the samples are used for the test set. Moreover, it applies the polynomial kernel of degree 4 for feature data and label data. The number of generations is set to 40, the size of sparrow population is set to 70, and the proportion of producers (PD), the proportion of scouts (SD) and the safety threshold (ST) are set to 0.2, 0.1, and 0.8, respectively.

In this section, hamming loss, one-error, coverage, ranking loss and average precision are selected to measure the performance of the above methods. Tables 3–7 show the classification performance of these multi-label feature selection methods. The boldface in the table indicates the best performance, and the numbers in parentheses indicate the relative ranking of the seven algorithms on each evaluation metric for each dataset.

Table 3. Comparison of hamming loss of different algorithms on each dataset.

dataset	MLSSA	FIMF	GRRO	GRRO-LS	PPT-CHI	PPT-MI	Ant-TD
medical	0.0228(3)	0.0214(2)	0.0271(4)	0.0271(4)	0.0280(6)	0.0281(7)	**0.0168(1)**
languagelog	**0.1805(1)**	0.1919(2)	0.2092(5)	0.2060(3)	0.2136(6)	0.2192(7)	0.2067(4)
tmc2007	**0.0924(1)**	0.0961(6)	0.0968(7)	0.0956(5)	0.0938(3)	0.0948(4)	0.0933(2)
corel5k	0.0052(2)	0.0063(3)	0.0102(6)	0.0102(6)	0.0097(4)	0.0097(4)	**0.0037(1)**
enron	**0.0329(1)**	0.0379(3)	0.0503(6)	0.0503(6)	0.0422(5)	0.0412(4)	0.0365(2)
chess	0.0086(2)	0.0110(5)	0.0109(3)	0.0109(3)	0.0113(7)	0.0111(6)	**0.0081(1)**
genbase	0.0245(2)	0.0252(3)	0.0434(6)	0.0441(7)	0.0263(4)	0.0351(5)	**0.0158(1)**
delicious	**0.0193(1)**	0.0198(5)	0.0198(5)	0.0197(4)	0.0195(2)	0.0196(3)	0.0198(5)

Table 4. Comparison of one-error of different algorithms on each dataset.

dataset	MLSSA	FIMF	GRRO	GRRO-LS	PPT-CHI	PPT-MI	AntTD
medical	0.4550(2)	0.4586(3)	0.9206(7)	0.9203(6)	0.7133(5)	0.6836(4)	**0.2695(1)**
languagelog	0.2142(2)	0.2444(3)	0.3186(5)	0.3195(6)	0.3208(7)	0.3935(4)	**0.1518(1)**
tmc2007	**0.4675(1)**	0.5219(7)	0.4750(3)	0.4872（5）	0.4836(4)	0.4906(6)	0.4748(2)
corel5k	0.0883(2)	0.1150(3)	0.5550(4)	0.6250（5）	0.6317(6)	0.6383(7)	**0.0280(1)**
enron	**0.2197(1)**	0.2681(3)	0.5167(6)	0.5167(6)	0.3611(4)	0.4094(5)	0.2221(2)
chess	0.4325(2)	0.7686(5)	0.9211(6)	0.9247(7)	0.7594(4)	0.7550(3)	**0.2163(1)**
genbase	0.2713(2)	0.3408(4)	0.3570(5)	0.4085(7)	0.2775(3)	0.3962(6)	**0.0825(1)**
delicious	0.6002(7)	0.5808(3)	0.5955(6)	0.5852(4)	**0.5660(1)**	0.5890(5)	0.5782(2)

It can be seen from Tables 3–7 that the MLSSA is superior or competitive to FIMF, GRRO, GRRO-LS, PPT-CHI, PPT-MI and Ant-TD algorithms in terms of hamming loss, one-error, coverage, ranking loss, and average precision. Specifically, it can be seen from Tables 3 and 6 that MLSSA

ranks first in four datasets among the seven algorithms, and ranks second or third on the other datasets. In Table 4, although the MLSSA ranks first only in two datasets, except for the last ranking on the delicious dataset, it ranks second in other datasets, which is relatively superior to other algorithms. In Table 5, the MLSSA ranks first in six datasets, and ranks second and third in other two datasets, which is significantly better than other comparison algorithms. In Table 7, although the proposed algorithm ranks first only in two datasets, it ranks second in most of the other datasets. Therefore, it can be seen that MLSSA can achieve better classification performance.

Table 5. Comparison of coverage of different algorithms on each dataset.

dataset	MLSSA	FIMF	GRRO	GRRO-LS	PPT-CHI	PPT-MI	Ant-TD
medical	4.5064(2)	5.4106(3)	9.8267(6)	9.8886(7)	7.0520(5)	6.9167(4)	**3.6900(1)**
languagelog	**51.0961(1)**	53.0903(4)	51.5669(2)	55.6447(6)	55.0808(5)	55.7092(7)	52.4532(3)
tmc2007	**5.3386(1)**	5.7258(3)	6.7047(7)	6.6597(6)	5.9786(4)	6.6472(5)	5.3622(2)
corel5k	**43.9908(1)**	57.0614(5)	54.7569(3)	55.3606(4)	58.5689(6)	63.4750(7)	50.9270(2)
enron	**9.8725(1)**	11.3911(4)	12.5458(7)	12.4744(6)	10.6347(2)	10.6708(3)	11.4839(5)
chess	**66.9128(1)**	77.1972(5)	82.4578(6)	83.0583(7)	75.9845(3)	76.9500(4)	72.1711(2)
genbase	1.6213(3)	2.1339(5)	3.5687(6)	3.9832(7)	1.4208(2)	1.7210(4)	**0.9275(1)**
delicious	**746.7223(1)**	749.4578(2)	758.6830(4)	757.8877(3)	760.5303(5)	763.1310(7)	762.1550(6)

Table 6. Comparison of ranking loss of different algorithms on each dataset.

dataset	MLSSA	FIMF	GRRO	GRRO-LS	PPT-CHI	PPT-MI	Ant-TD
medical	0.0891(2)	0.1118(3)	0.2088(6)	0.2104(7)	0.1434(5)	0.1379(4)	**0.0674(1)**
languagelog	0.2119(2)	0.2261(3)	0.2523(6)	0.2590(7)	0.2429(5)	0.2350(4)	**0.2046(1)**
tmc2007	**0.1345(1)**	0.1514(3)	0.1799(7)	0.1789(6)	0.1515(4)	0.1759(5)	0.1369(2)
corel5k	**0.0378(1)**	0.0529(3)	0.0655(6)	0.0673(7)	0.0564(4)	0.0642(5)	0.0463(2)
enron	**0.0784(1)**	0.0925(5)	0.1148(7)	0.1101(6)	0.0880(2)	0.0899(3)	0.0911(4)
chess	**0.1758(1)**	0.2014(3)	0.2456(6)	0.2478(7)	0.2153(4)	0.2181(5)	0.1780(2)
genbase	0.0463(3)	0.0618(5)	0.1150(7)	0.1301(6)	0.0390(2)	0.0528(4)	**0.0165(1)**
delicious	0.2131(3)	0.2174(7)	0.2168(5)	0.2168(5)	**0.2050(1)**	0.2070(2)	0.2161(4)

Table 7. Comparison of average precision of different algorithms on each dataset.

dataset	MLSSA	FIMF	GRRO	GRRO-LS	PPT-CHI	PPT-MI	Ant-TD
medical	0.6347(2)	0.6157(3)	0.2209(6)	0.2184(7)	0.3390(5)	0.4165(4)	**0.7762(1)**
languagelog	0.6087(2)	0.5823(3)	0.5268(6)	0.5112(7)	0.5289(5)	0.5447(4)	**0.6265(1)**
tmc2007	**0.6045(1)**	0.5570(3)	0.5448(6)	0.5499(5)	0.5591(3)	0.5381(7)	0.6017(2)
corel5k	0.7663(2)	0.6892(3)	0.3466(6)	0.3178(7)	0.3701(4)	0.3701(4)	**0.8379(1)**
enron	**0.7462(1)**	0.7035(3)	0.5672(7)	0.5783(6)	0.6842(4)	0.6825(5)	0.7197(2)
chess	0.4460(2)	0.2357(4)	0.1216(6)	0.1189(7)	0.2419(3)	0.2349(5)	**0.5499(1)**
genbase	0.8119(2)	0.7532(4)	0.6886(6)	0.6460(7)	0.8116(3)	0.7265(5)	**0.9379(1)**
delicious	0.2083(3)	0.2068(4)	0.1993(7)	0.1994(6)	0.2128(2)	**0.2156(1)**	0.2066(5)

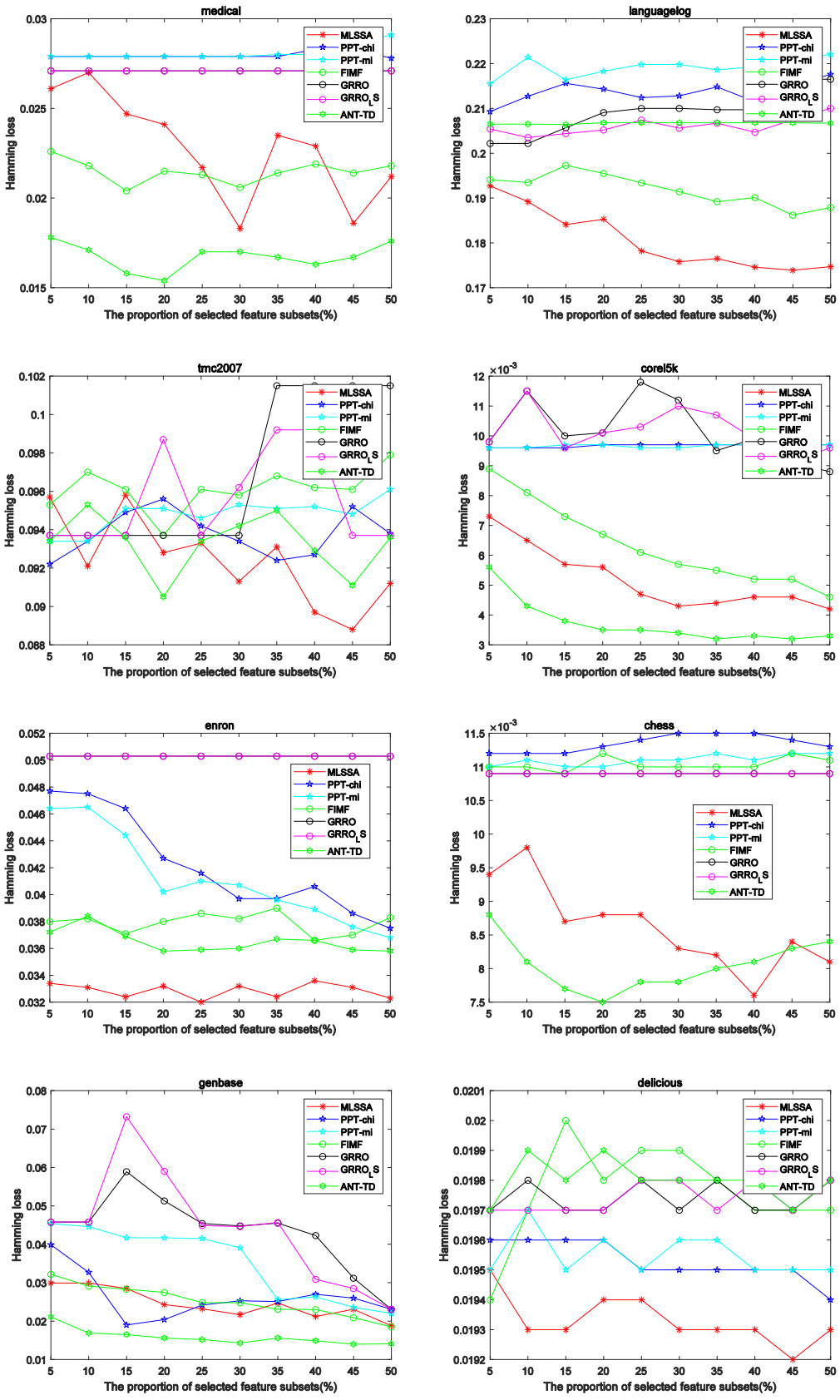

Figure 2. Comparison of hamming loss of different algorithms.

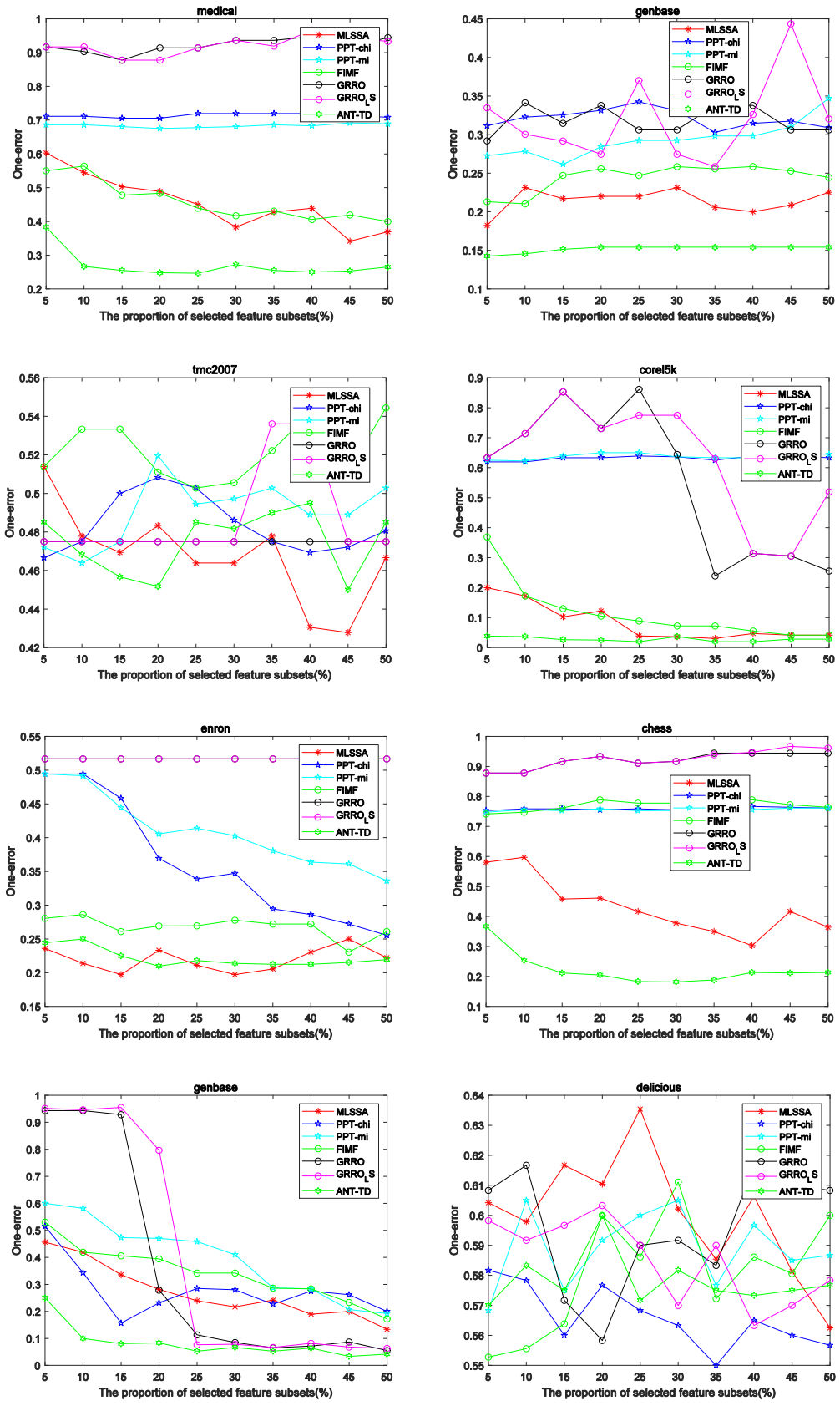

Figure 3. Comparison of one-error of different algorithms.

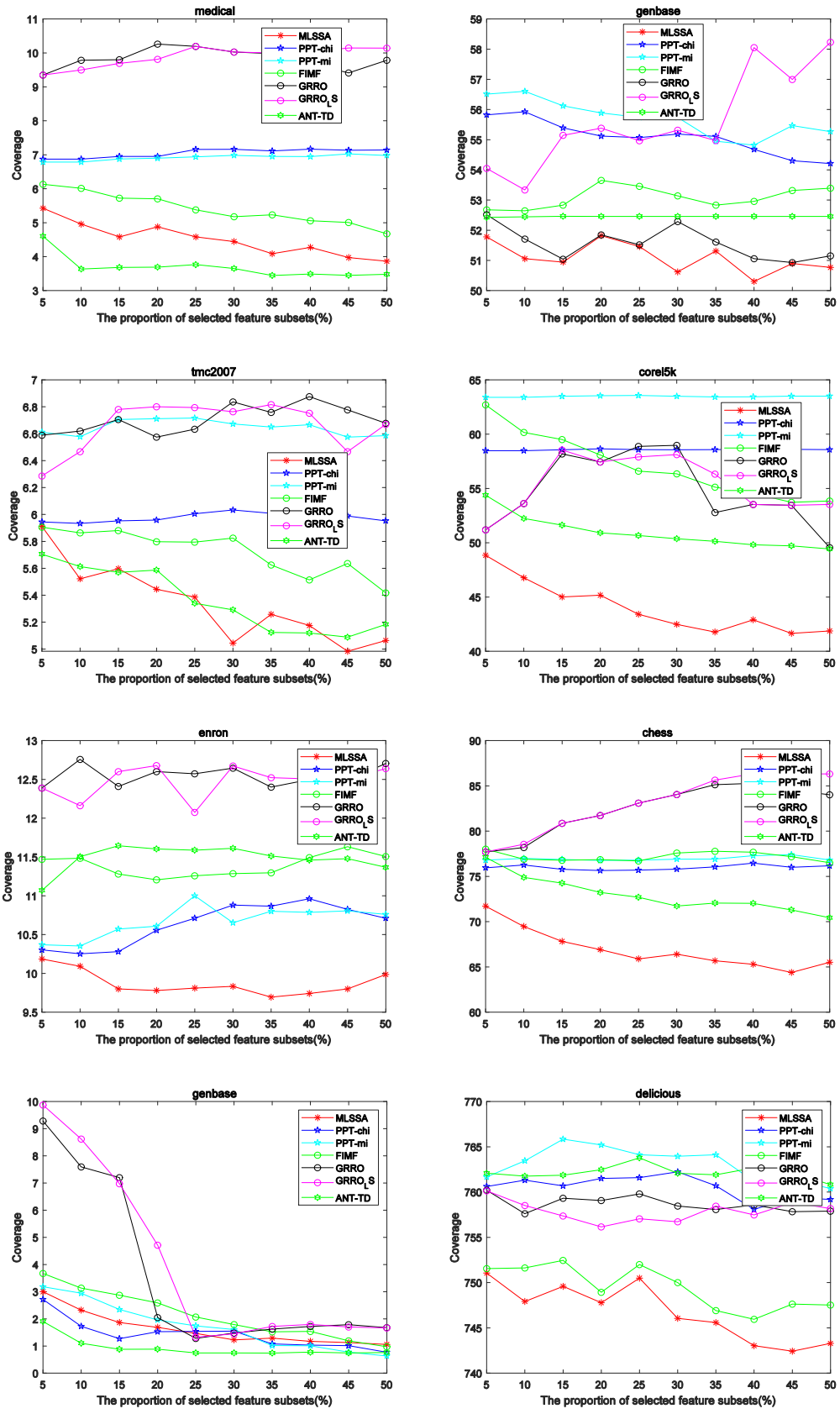

Figure 4. Comparison of coverage of different algorithms.

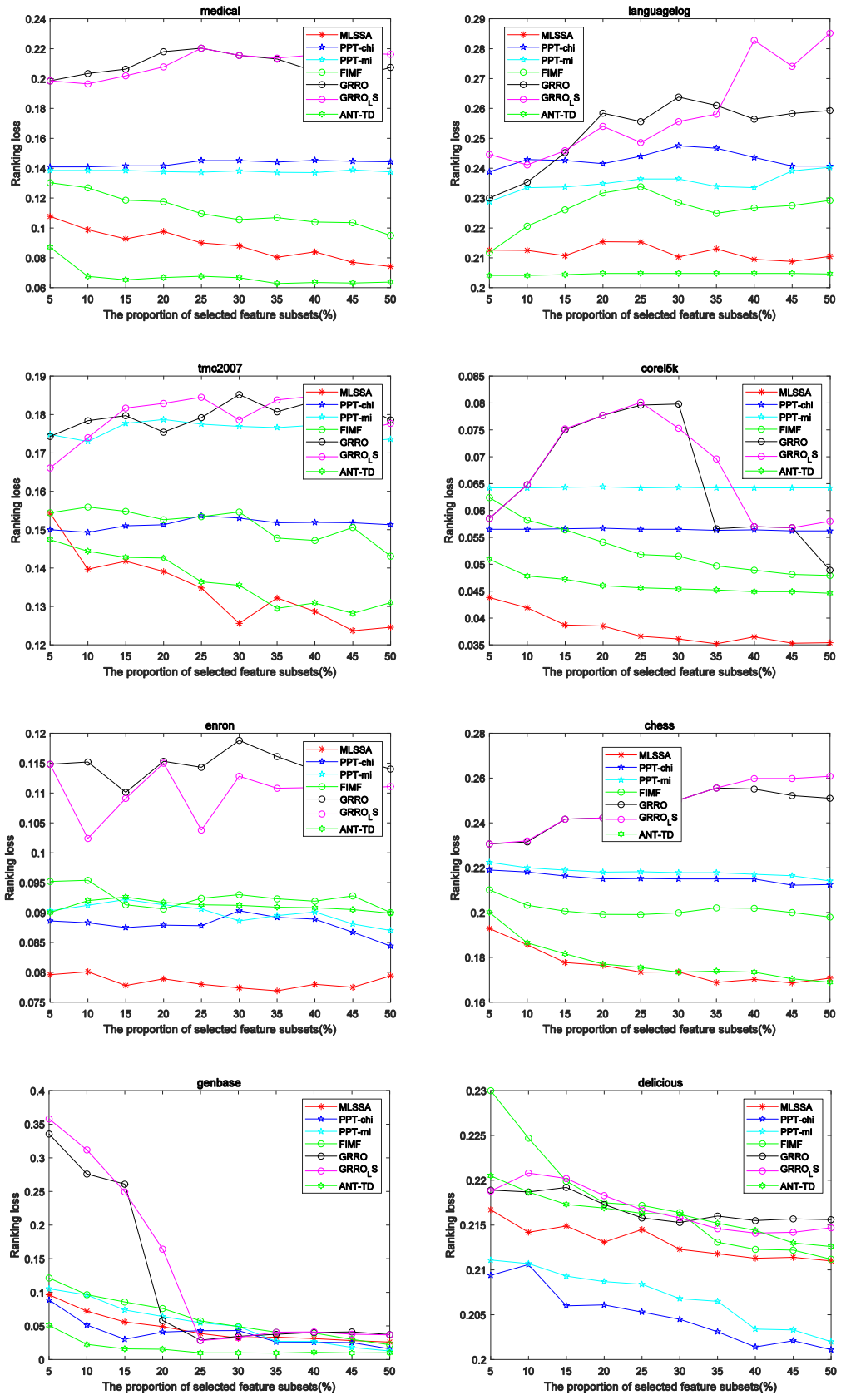

Figure 5. Comparison of ranking loss of different algorithms.

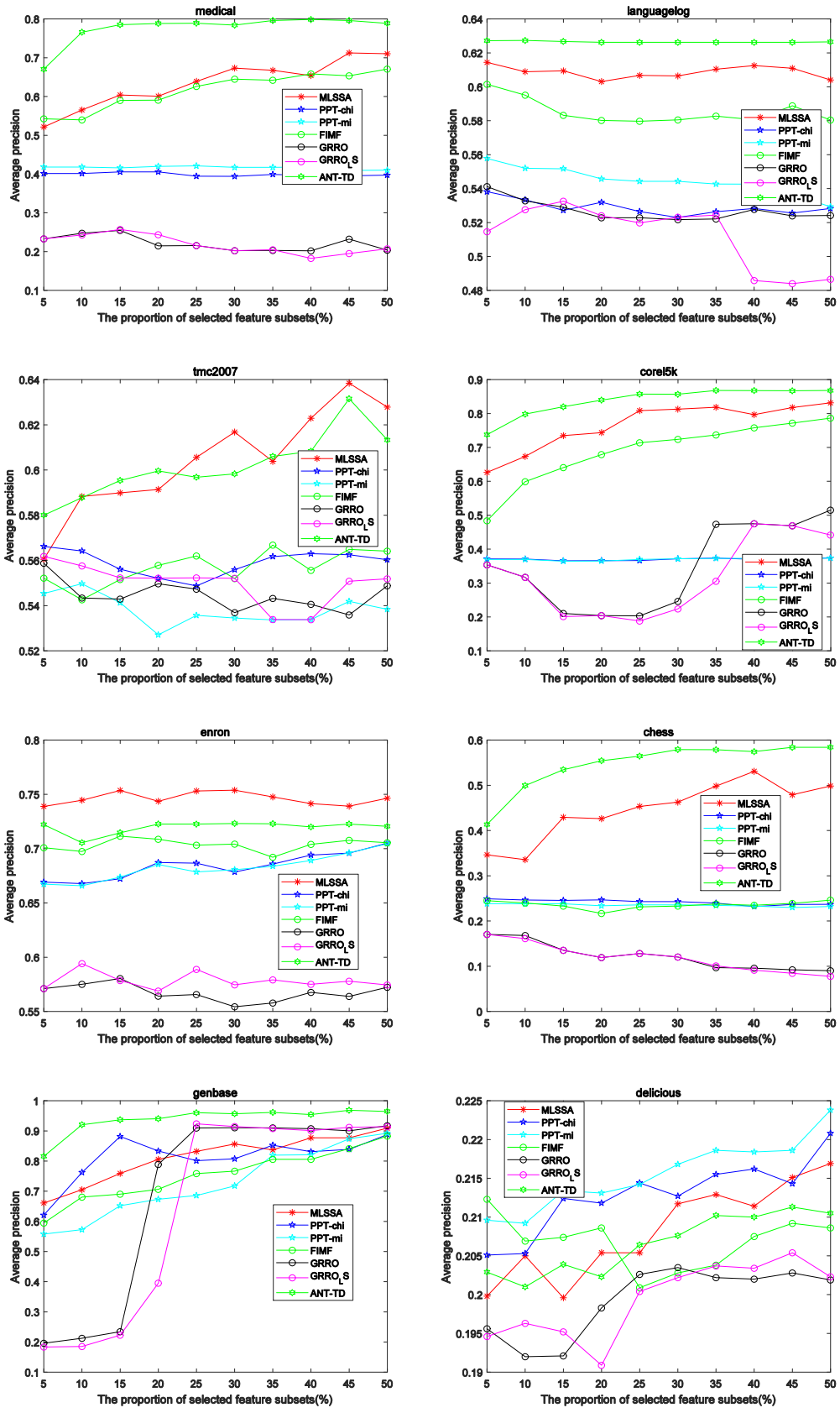

Figure 6. Comparison of average precision of different algorithms.

In order to clearly show the classification performance of MLSSA and other comparison algorithms, Figures 2–6 show the performance of these feature selection algorithms. The horizontal axis represents the number of features selected by each feature selection algorithm, and the vertical axis represents the value obtained by each algorithm according to the evaluation indicators. The results show that the proposed MLSSA performs better than the comparison algorithm in most cases. Specifically, in Figures 2–5, the curve of MLSSA is generally below the curves of other algorithms. In Figure 6, the curve of MLSSA is generally above the curves of other algorithms.

5. Conclusions

This paper proposes a new multi-label feature selection algorithm based on HSIC and SSA (MLSSA). By utilizing the HSIC as a feature selection criterion to describe the dependency between features and all labels, MLSSA attempts to search in the feature space to find the optimal features. The performance of the proposed method is compared with those of FIMF, GRRO, GRRO-LS, PPT-CHI, PPT-MI and Ant-TD algorithms on eight datasets. Experimental results demonstrate the effectiveness of the method. However, the algorithm does not take into account the correlation between labels and the problem that the SSA algorithm can easily fall into local optima. Therefore, in future research, these two types of problems should be further studied to further improve the performance of the algorithm. Last but not least, more comparison experiments with the state-of-the-art multi-label feature selection methods on more real-world datasets should be further investigated to verify the effectiveness of the proposed method.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This paper is supported in part by the National Natural Science Foundation of China (No. 61966002) and the Science and Technology Program Foundation of Jiangxi Education Committee of China (No. GJJ2201203).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, et al., Feature selection: a data perspective, *ACM Comput. Surv.*, **50** (2018), 1–45. https://doi.org/10.1145/3136625
2. H. Zhou, T. Wang, D. Zhang, Research progress of multi-label feature selection, *Comput. Eng. Appl.*, **58** (2022), 52–67. https://doi.org/10.3778/J.ISSN.1002-8331.2202-0114

3. T. Wang, X. Dai, Y. Liu, Learning with Hilbert-Schmidt independence criterion: A review and new perspectives, *Knowl. Based Syst.*, **234** (2021), 107567. https://doi.org/10.1016/j.knosys.2021.107567
4. A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, et al., A review of clustering techniques and developments, *Neurocomputing*, **267** (2017), 664–681. https://doi.org/10.1016/j.neucom.2017.06.053
5. S. Ayesha, M. K. Hanif, R. Talib, Overview and comparative study of dimensionality reduction techniques for high dimensional data, *Inf. Fusion*, **59** (2020), 44–58. https://doi.org/10.1016/j.inffus.2020.01.005
6. T. Wang, Z. Hu, H. Liu, A unified view of feature selection based on Hilbert-Schmidt independence criterion, *Chem. Intell. Lab. Syst.*, **236** (2023), 104807. https://doi.org/10.1016/j.chemolab.2023.104807
7. A. Tharwat, Independent component analysis: An introduction, *Appl. Comput. Inf.*, **17** (2021), 222–249. https://doi.org/10.1016/j.aci.2018.08.006
8. Y. Zhang, X. Xiu, Y. Yang, W. Liu, Fault detection based on canonical correlation analysis with rank constrained optimization, in *The 2021 40th Chinese Control Conference*, (2021). https://doi.org/10.26914/c.cnkihy.2021.028664
9. L. Zhang, T. Wang, H. Zhou, A multi-strategy improved sparrow search algorithm, *Comput. Eng. Appl.*, **58** (2022), 133–140. https://doi.org/10.3778/j.issn.1002-8331.2112-0427
10. M. Paniri, M. B. Dowlatshahi, H. Nezamabadi-pour, MLACO: A multi-label feature selection algorithm based on ant colony optimization, *Knowl. Based Syst.*, **193** (2019), 105285. https://doi.org/10.1016/j.knosys.2019.105285
11. M. Paniri, M. B. Dowlatshahi, H. Nezamabadi-pour, Ant-TD: Ant colony optimization plus temporal difference reinforcement learning for multi-label feature selection, *Swarm Evol. Comput.*, **64** (2021), 100892. https://doi.org/10.1016/j.swevo.2021.100892
12. Y. Zhang, D. Gong, X. Sun, Y. Guo, A PSO-based multi- objective multi-label feature selection method in classification, *Sci. Rep.*, **7** (2017), 376. https://doi.org/10.1038/s41598-017-00416-0
13. D. Paul, A. Jain, S. Saha, J. Mathew, Multi-objective PSO based online feature selection for multi-label classification, *Knowl. Based Syst.*, **222** (2022), 106966. https://doi.org/10.1016/j.knosys.2021.106966
14. Z. Lu, X. Cheng, Y. Zhang, Global optimization method based on consensus particle swarm, *J. Syst. Simul.*, **32** (2020), 1936–1942. https://doi.org/10.16182/j.issn1004731x.joss.20-fz0371
15. M. Abdel-Basset, D. El-Shahat, I. El-Henawy, V. Albuquerque, S. Mirjalili, A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection, *Expert Syst. Appl.*, **139** (2020), 112824. https://doi.org/10.1016/j.eswa.2019.112824
16. W. Li, Y. Li, Y. Zhao, B. Yan, Research on particle filter algorithm based on improved grey wolf algorithm, *J. Syst. Simul.*, **33** (2021), 37–45. https://doi.org/10.16182/j.issn1004731x.joss.19-0276
17. J. Xue, B. Shen, A novel swarm intelligence optimization approach: sparrow search algorithm, *Syst. Sci. Control Eng.*, **8** (2020), 22–34. https://doi.org/10.1080/21642583.2019.1708830
18. L. Sun, Y. Chen, J. Xu, Multi-label feature selection algorithm based on improved ReliefF, *J. Shandong Univ. Nat. Sci.*, **57** (2022), 1–11. https://doi.org/10.6040/j.issn.1671-9352.7.2021.167

19. J. Gonzalez-Lopez, S. Ventura, A. Cano, Distributed multi-label feature selection using individual mutual information measures, *Knowl. Based Syst*., **188** (2020), 105052. https://doi.org/10.1016/j.knosys.2019.105052
20. J. Gonzalez-Lopez, S. Ventura, A. Cano, Distributed selection of continuous features in multilabel classification using mutual information, *IEEE Trans. Neural Networks Learn. Syst*., **31** (2020), 2280–2293. https://doi.org/10.1109/TNNLS.2019.2944298
21. C. Xiong, W. Qian, Y. Wang, J. Huang, Feature selection based on label distribution and fuzzy mutual information, *Inf. Sci*., **574** (2021), 297–319. https://doi.org/10.1016/j.ins.2021.06.005
22. Z. Sha, Z. Liu, C. Ma, J Chen, Feature selection for multi-label classification by maximizing full-dimensional conditional mutual information, *Appl. Intell.,* **51** (2021), 326–340. https://doi.org/10.1007/s10489-020-01822-0
23. C. Liu, Q. Ma, J. Xu, Multi-label feature selection method combining unbiased Hilbert-Schmidt independence criterion with controlled genetic algorithm, *Lect. Notes Comput. Sci.*, **11304** (2018), 3–14. https://doi.org/10.1007/978-3-030-04212-7_1
24. G. Li, Y. Li, Y. Zheng, Y. Li, Y. Hong, X. Zhou, A novel feature selection approach with Pareto optimality for multi-label data. *Appl. Intell*., **51** (2021), 7794–7811. https://doi.org/10.1007/s10489-021-02228-2
25. G. Li, Y. Li, Y. Zheng, A novel multi-label feature selection based on pareto optimality, *Lect. Notes Data Eng. Commun. Technol*., **88** (2021), 1010–1016. https://doi.org/10.1007/978-3-030-70665-4_109
26. Y. Li, *Binary sparrow search algorithm and its application in feature selection*, Master thesis, Tianjin Normal University, 2022. https://doi.org/10.27363/d.cnki.gtsfu.2022.000316
27. T. Wang, W. Li, Kernel learning and optimization with Hilbert-Schmidt independence criterion, *Int. J. Mach. Learn. Cybern*., **9** (2018), 1707–1717. https://doi.org/10.1007/s13042-017-0675-7
28. Z. Hu, T. Wang, H. Zhou, Review of feature selection methods based on kernel statistical independence criteria, *Comput. Eng. Appl*., **58** (2022), 54–64. https://doi.org/10.3778/j.issn.1002-8331.2203-0527
29. X. Tian, J. He, Y. Shi, Statistical dependence test with Hilbert-Schmidt independence criterion, *J. Phys. Confer. Ser*., **1601** (2020), 032008. https://doi.org/10.1088/1742-6596/1601/3/032008
30. B. B. Damodaran, N. Courty, S. Lefèvre, Sparse Hilbert Schmidt independence criterion and surrogate-kernel-based feature selection for hyperspectral image classification, *IEEE Trans. Geosci. Remote Sens*., **55** (2017), 2385–2398. https://doi.org/10.1109/TGRS.2016.2642479
31. X. Lü, X. Mu, J. Zhang, Z. Wang, Chaotic sparrow search optimization algorithm, *J. Beijing Univ. Aeronaut. Astronaut*., **47** (2021), 1712–1720. https://doi.org/10.13700/j.bh.1001-5965.2020.0298
32. M. L. Zhang, Z. H. Zhou, A review on multi-label learning algorithms, *IEEE Trans. Knowl. Data Eng*., **26** (2014), 1819–1837. https://doi.org/10.1109/TKDE.2013.39
33. J. Zhang, Y. Lin, M. Jiang, S. Li, Y. Tang, K. C. Tan, Multi-label feature selection via global relevance and redundancy optimization, in *The 29th International Joint Conference on Artificial Intelligence*, (2020). https://doi.org/10.24963/ijcai.2020/348
34. J. Lee, D. W. Kim, Fast multi-label feature selection based on information-theoretic feature ranking, *Pattern Recognit.*, **48** (2015), 2761–2771. https://doi.org/10.1016/j.patcog.2015.04.009

35. G. Doquire, M. Verleysen, Mutual information-based feature selection for multilabel classification, *Neurocomputing*, **122** (2013), 148–155. https://doi.org/10.1016/j.neucom.2013.06.035
36. G. Doquire, M. Verleysen, Feature selection for multi-label classification problems, in *The 11th International Conference on Artificial Neural Networks*, (2011). https://doi.org/10.1007/978-3-642-21501-8_2
37. K. Trochidis, G. Tsoumakas, G. Kalliris, I. Vlahavas, Multilabel classification of music into emotions, in *The 9th International Conference on Music Information Retrieval*, (2008). https://doi.org/10.1186/1687-4722-2011-426793

AIMS Press