A Multi-label Filter Feature Selection Method Based on Approximate Pareto Dominance

Jian Zhou*, Yinnong Guo
School of Management Engineering, Qingdao University of Technology, Qingdao, China

***Abstract*—The Pareto dominance has been applied to resolve the issue of choosing significant features from a multi-label dataset. High-dimensional labels will directly result in the difficulty of forming Pareto dominance. This work proposes a multi-label feature selection approach based on the approximate Pareto dominance (MAPD) to address this issue. It maps the multi-label feature selection to the problem of solving the approximate Pareto dominant solution set. By introducing an approximate parameter, it is possible to efficiently cut down on the amount of features in the chosen feature subset while also raising its quality. To verify the performance of MAPD, this research compares the MAPD algorithm with alternative approaches in terms of Hamming loss, accuracy, and chosen feature size using nine publicly available multi-label datasets. The findings indicate that the MAPD method performs better in terms of classification accuracy, Hamming loss, and the amount of features that may be chosen.**

Keywords*—Approximate Pareto dominance; multi-label data; feature selection

I. Introduction

Feature selection is a process of removing noisy information and selecting the most significant feature subset, which is commonly considered as a pre-process of building a classifier machine learning model [1]-[3]. The multi-label feature selection problem is more universal in application than the single-label feature selection problem [4], [5]. For example, it might be necessary to simultaneously judge the geographical location, weather conditions, and image content of a figure in the process of image recognition [6], [7]. When processing the text categorization, we may need to judge whether the text belongs to multiple bibliographic categories [8], [9]. It also might be necessary to judge whether a protein has multiple different functions in the field of bioinformatics in the same manner [10].

The multi-label feature selection methods could be classified as the filter methods [11]-[13], the wrapper methods [14], [15], and the embedded methods [16], [17]. Since the increase of label dimension would lead to higher time complexity of the feature selection processing, this paper only focuses on the filter feature selection methods that are more efficient compared with the wrapper and embedded methods.

In the current literature, the multi-label feature selection problem can be resolved primarily in two ways. The first strategy is to convert the multi-label data into single-label data and then choose feature subsets using single-label feature selection techniques [18], [19]. However, such methods create an abundance of labels with only a limited number of observations which is not beneficial for establishing a classifier model. In order to improve the disadvantages of such methods, a method called pruned problem transformation (PPT) is proposed, and it ignores the labels with observations lower than the given threshold [20]. This method can ensure that each converted single-label has enough observations to establish a classification model, but this irreversible conversion may lose some label information [21]. The second method involves choosing a feature subset using a specific multi-label feature selection algorithm [22]-[25]. For example, an approximating mutual information (AMI) method is proposed, and it uses the feature selection criterion as maximizing the mutual information between features and labels and minimizing the mutual information among features [26]. A multi-label feature selection strategy based on a scalable criterion for a large label collection (SCLS) is proposed, which can evaluate the conditional correlation between variables more accurately through an extensible correlation evaluation process [27]. Due to the information loss issue of the first way, it is believed that the second way has better performance under several evaluation criteria, such as classification accuracy and Hamming loss [28].

Recently, scholars have applied the concept of Pareto dominance to multi-label feature selection problems. For resolving the multi-label feature selection problem, the Pareto dominance concept is appropriate, since it can transform this problem into a more manageable issue. Specifically, a multi-label feature selection technique based on the Pareto dominance concept (ParFS) is proposed [29]. The ParFS algorithm treats each label as a dimension of Pareto dominance, and thus the issue of multi-label feature selection becomes a Pareto dominance problem. The key point of this algorithm is that the original feature set is viewed as a solution set. Each feature in the original feature set is viewed as a solution, and the evaluation function of the solution is regarded as a feature's and a label's correlation vector. Then the feature selection problem is transformed into how to delete those non-Pareto optimal solutions. In fact, high-dimensional data refers not only to the high-dimensional features of the data, but also to its high-dimensional labels. High-dimensional labels increase the difficulty of using the ParFS algorithm to resolve issues of multi-label feature selection [30], [31]. Specifically, the increase of label dimensions will directly lead to the increase of the dimensions to be considered in Pareto dominance, which results in the difficulty of forming Pareto dominance, causing the Pareto-dominance-based algorithm to fail to finish the multi-label feature selection task. Consequently, it's essential to improve the multi-label feature selection method based on Pareto dominance concept.

In this paper, a multi-label feature selection strategy based on approximate Pareto dominance is proposed (MAPD). Approximate Pareto dominance requires that one solution is superior to the others in most dimensions, but not in all dimensions. Compared with the existing concept of Pareto dominance, the concept of approximate Pareto dominance introduces a new approximate parameter. This parameter can reduce the difficulty of forming approximate Pareto dominance between two solutions under the evaluation function of high-dimensional solutions, and ensure that the scale of approximate Pareto dominance solutions is within an acceptable range. The three main contributions are as follows:

1) A new concept called approximate Pareto dominance is proposed. By introducing an approximate parameter, it can solve the problem when Pareto dominance is difficult to form in the case that the evaluation function dimension of the solution is high.

2) Approximate Pareto dominance is applied in the multi-label feature selection, and the multi-label feature selection issue is mapped to the challenge of determining the approximate Pareto dominance solution set.

3) Based on the approximate Pareto dominance, MAPD is built for the high-dimensional multi-label feature selection problem, and it is proved to be competitive compared with the existing methods.

The remainder of the paper is organized as follows: Preliminaries of this work are presented in Section II. The proposed approximate Pareto dominance concept and multi-label feature selection method is discussed in Section III. Experimental studies and discussion are presented in Section IV, and the paper is concluded in Section V.

II. PRELIMINARIES

A. Problem Description

Given a dataset, $\boldsymbol{X}=[\boldsymbol{X}_1,...,\boldsymbol{X}_m]$ denotes the feature observation space, and its corresponding *l*-dimension label space is $\boldsymbol{Y}=[\boldsymbol{Y}_1,...,\boldsymbol{Y}_l]$, where $\boldsymbol{X}_j=[x_{1j},\cdots,x_{ij},\cdots,x_{nj}]^T$ is the *j*th feature in $\boldsymbol{X}$, x_{ij} denotes the *i*th observation of the *j*th feature, $\boldsymbol{Y}_t=[y_{1t},\cdots,y_{it},\cdots,y_{nt}]^T$ is the *t*th label in $\boldsymbol{Y}$, y_{it} is the *i*th observation of the *t*th label, $y_{it}=\{0\}\ or\ \{1\}$, $i=1,2,...,n$, $j=1,2,...,m$, and $t=1,2,...,l$. To create the classification machine learning model, the multi-label feature selection in this study aims to identify the best feature subset from all feasible subsets.

B. Symmetrical Uncertainty

Symmetrical Uncertainty which abbreviated to SU is a measure of the degree to which two variables are related [32]. In essence, SU measures the information that the two variables exchange and is a standardized representation of the mutual information. In other words, SU quantifies how much one variable's uncertainty is reduced when the other variable is known, and the higher the degree of SU, the more knowledge the two variables have in common. The formula for calculating SU is given below [33].

$$SU(\boldsymbol{U},\boldsymbol{V})=2\frac{H(\boldsymbol{U})-H(\boldsymbol{U}|\boldsymbol{V})}{H(\boldsymbol{U})+H(\boldsymbol{V})}, \quad (1)$$

$$H(\boldsymbol{U})=-\sum_{i=1}^{n}p(u_i)\log_2(p(u_i)), \quad (2)$$

$$H(\boldsymbol{U}|\boldsymbol{V})=-\sum_{j=1}^{n}p(v_j)\sum_{i=1}^{n}p(u_i|v_j)\log_2(p(u_i|v_j)), \quad (3)$$

In the above equation, $\boldsymbol{U}$ and $\boldsymbol{V}$ denote two variables with n observations. $H(\boldsymbol{U})$ and $H(\boldsymbol{V})$ are respectively the entropy of $\boldsymbol{U}$ and the entropy of $\boldsymbol{V}$. $H(\boldsymbol{U}|\boldsymbol{V})$ is the conditional entropy of $\boldsymbol{U}$ under $\boldsymbol{V}$.

C. Pareto Dominance

The following definition of Pareto dominance is used to compare the results of two solutions to a particular problem.

Definition 1 (Pareto Dominance [34]): If $s_1(s_{11},s_{12},...,s_{1n})$ and $s_2(s_{21},s_{22},...,s_{2n})$ are two solutions of a given problem, and $g(s_i)=(g_1(s_i),g_2(s_i),...,g_m(s_i)), i\in\{1,2\}$ is the evaluation function of m dimensions of the given problem,

1) we define that the solution $s_2(s_{21},s_{22},...,s_{2n})$ is Pareto dominated to the solution $s_1(s_{11},s_{12},...,s_{1n})$ if and only if $g_j(s_1)>g_j(s_2), j\in\{1,2,...,m\}$ is satisfied;

2) we define that the solution $s_2(s_{21},s_{22},...,s_{2n})$ is weakly Pareto dominated to the solution $s_1(s_{11},s_{12},...,s_{1n})$ if and only if $g_j(s_1)\geq g_j(s_2), j\in\{1,2,...,m\}$ is satisfied;

3) we define that the solution $s_1(s_{11},s_{12},...,s_{1n})$ and solution $s_2(s_{21},s_{22},...,s_{2n})$ have no differences under the Pareto dominance if and only if $g_j(s_1)>g_j(s_2)$ and $g_k(s_1)<g_k(s_2),\ j\neq k\in\{1,2,...,m\}$ are satisfied.

It can be inferred that the conditions required for Pareto dominance are relatively strict, which requires that one solution is superior to another solution in each dimension of the evaluation function. Considering the increased evaluation function dimensions of the solution, we assume that all the evaluation function dimensions of the solutions are independent with each other, and the difficulty of forming Pareto dominance will increase exponentially. Therefore, for the high-dimensional evaluation function of the solutions, one solution is hard to be Pareto dominated to another solution.

We concentrate on the Pareto dominance relationships between one solution and other solutions when there are more than two possible solutions to the problem. The set of Pareto optimal solutions is defined as follows.

Definition 2 (Pareto Optimal Solutions Set [34]): If $S=\{s_1,s_2,...,s_n\}$ is a solutions set of a specific problem and $s_l\in S$ is not Pareto dominated to any other solutions in S, then we propose that $s_l\in S$ is a Pareto optimal solution. All the Pareto optimal solutions in S is called as the set of Pareto optimal solutions.

According to Definition 2, we suggest that the set of Pareto-optimal solutions can partially substitute for the set of

original solutions as the remaining solutions are all inferior to a certain solution in the Pareto optimal solutions set.

III. Approximate Pareto Dominance and Multi-Label Feature Selection

A. Multi-label Feature Selection Based on Pareto Dominance

Given a multi-label dataset with *m* features, *l* labels and *n* observations, the original features are denoted as a set of solutions of the feature selection problem, and each feature in the original feature set is denoted as a solution of the feature selection problem. The *l*-dimensional evaluation function of the solution is defined as the symmetric uncertainty between the feature and its *l* labels. In this way, choosing a feature subset from the original feature set is analogous to choosing the Pareto optimal solutions set from the initial solution set.

As mentioned in Definition 1, Pareto dominance strictly requires that one solution is superior to another solution in each dimension of the evaluation function of the solution. In particular, with the increased dimensions of the evaluation function of a solution, it is challenging to come up with a solution that is superior to another solution in all the dimensions of the evaluation function. As such, most solutions in the solution set become Pareto optimal solutions. In this case, most features in the original feature set are preserved in the feature selection process. Pareto dominance might cause the inefficiencies of the feature selection process because useless features cannot be removed. Therefore, to cope with the multi-label feature selection issue, this study proposes the approximate Pareto dominance based on the Pareto dominance to avoid the above circumstance.

B. Approximate Pareto Dominance

Based on Pareto Dominance, this study proposes approximate Pareto dominance and the set of approximate Pareto optimal solutions.

Definition 3 (Approximate Pareto Dominance): If $s_1(s_{11}, s_{12}, ..., s_{1n})$ and $s_2(s_{21}, s_{22}, ..., s_{2n})$ are two solutions of a problem, and $g(s_i) = (g_1(s_i), g_2(s_i), ..., g_m(s_i)), i \in \{1, 2\}$ is the evaluation function of m dimensions of the given problem,

1) we define solution $s_2(s_{21}, s_{22}, ..., s_{2n})$ is approximate Pareto dominated to solution $s_1(s_{11}, s_{12}, ..., s_{1n})$ if and only if $\sum_{j=1}^{m} \delta(g_j(s_1) > g_j(s_2)) > \alpha m$ is satisfied;

2) we define solution $s_2(s_{21}, s_{22}, ..., s_{2n})$ is weakly approximate Pareto dominated to solution $s_1(s_{11}, s_{12}, ..., s_{1n})$ if and only if $\sum_{j=1}^{m} \delta(g_j(s_1) > g_j(s_2)) \geq \alpha m$ is satisfied;

3) we define solution $s_1(s_{11}, s_{12}, ..., s_{1n})$ and solution $s_2(s_{21}, s_{22}, ..., s_{2n})$ have no differences under approximate Pareto dominance if and only if $\sum_{j=1}^{m} \delta(g_j(s_1) > g_j(s_2)) < \alpha m$ and $\sum_{j=1}^{m} \delta(g_j(s_2) > g_j(s_1)) < \alpha m$, $j \neq k \in \{1, 2, ..., m\}$ are satisfied.

Note that $\delta(x)$ is a conditional discriminant function and $0.5 < \alpha < 1$ is the approximate parameter. When condition x is satisfied, $\delta(x) = 1$; otherwise $\delta(x) = 0$.

Definition 4 (Approximate Pareto Optimal Solutions Set): If $S = \{s_1, s_2, ..., s_n\}$ is a set of solution of a given problem and $s_l \in S$ is not approximate Pareto dominated to any other solutions in S, then we define that $s_l \in S$ is an approximate Pareto optimal solution of the given problem. All the approximate Pareto optimal solutions in S are called the approximate Pareto optimal solutions set.

Compared with Pareto dominance, the approximate Pareto dominance introduces an approximate parameter. The higher the value of the approximate parameter is, the more dimensions of one solution are required to be superior to another solution in all the evaluation function dimensions, and the closer the approximate Pareto dominance is similar to the Pareto dominance. The upper bound of the approximate parameter is 1. When the approximate parameter approaches to the upper bound, the approximate Pareto dominance is equal to Pareto dominance. The lower bound of the approximate parameter is 0.5, which means that when one solution is approximate Pareto dominated to another solution, it is dominant in more than half of the dimensions. This approximate parameter can avoid the situation that two solutions are mutually dominant.

C. Multi-Label Feature Selection Algorithm

We build a multi-label feature selection algorithm (MAPD) based on approximate Pareto dominance for feature selection with high-dimensional labels, which is shown in the following Algorithm 1, where the SU_{ij} is the symmetrical uncertainty of the *i*th feature and the *j*th feature; the set *S* is the selected feature subset; $\delta(\cdot)$ is the conditional discriminant function; s_j is the *j*th solution in the original solution set (namely the *j*th feature); $g_p(s_j)$ is the *p*th dimension of the evaluation function of the *j*th solution; α is the approximate parameter. Approximate parameter α can keep the number of approximate Pareto dominance solutions in an acceptable range.

Algorithm 1 contains three important steps: (1) calculating the symmetrical uncertainty between one feature and one label (see lines 1-5 of Algorithm 1); (2) initializing the selected feature set as an empty set, treating each feature as a solution and the symmetric uncertainty between the feature and each label as one dimension in the evaluation function. We detect the approximate Pareto dominant relationship between each two features. If one feature is the approximate Pareto dominant solution of the original solution set, the feature is merged into the selected feature set until all features are checked (see lines 6-17 of Algorithm 1). (3) Output the selected feature set (see line 18 of Algorithm 1).

Algorithm 1 MAPD

Input: The approximate parameter α, the dataset $\boldsymbol{X}$ with m features, l labels and n observations;
Output: The finally chosen feature subset S.

1. *for* $i = 1:l$
2. *for* $j = 1:m$
3. $SU_{ij} = SU(\boldsymbol{X}_i, \boldsymbol{X}_j)$
4. *end*
5. *end*
6. $S = \varnothing$
7. *for* $i = 1:m$
8. $k = 0$
9. *for* $j = 1:m$
10. *if* $j \neq i \cap \sum_{p=1}^{l} \delta(g_p(s_j) > g_p(s_i)) > \alpha l$
11. $k = k + 1$
12. *end*
12. *end*
14. *if* $k = 0$
15. $S = S \cup i$
16. *end*
17. *end*
18. Output S

IV. Experimental Studies and Discussion

A. Parameter Setting

To evaluate the effectiveness of the suggested MAPD technique, we employ nine publicly available multi-label datasets (http://mulan.sourceforge.net/datasets-mlc.html). The datasets have been used in many studies, e.g., [35]-[37].

Given a dataset with m features, l labels and n observations, $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{il})$ is the real label of the ith observation, $\boldsymbol{y}'_i = (y'_{i1}, y'_{i2}, ..., y'_{il})$ is the predicted label of the ith observation based on the classification model. To assess how well multi-label feature selection approaches work in terms of precision, we use two criteria of Hamming loss and accuracy, which can be calculated as follows [38], [39]:

$$Hamming\ loss = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{m}\delta(y_{ij} \neq y'_{ij})}{l} \quad (4)$$

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta(\sum_{j=1}^{m}\delta(y_{ij} = y'_{ij}) = m)}{l} \quad (5)$$

According to Equation (4), Hamming loss analyzes each dimension of an observation's real label and prediction label and focuses on the prediction accuracy of a certain dimension. Then Hamming loss calculates the average error prediction rate of all observations in all dimensions. According to Equation (5), accuracy strictly requires that the real label and the prediction label should be exactly the same, and calculates the accuracy prediction based on all observations.

This study uses the average values of the accuracy and the Hamming loss of 5-fold cross validation of five times as a measure of how well feature selection techniques operate. Specifically, 5-fold cross validation means that the observations are randomly divided into five subsets. For a total of five tests, one of which is chosen as the testing set and the other four as the training set, and the average performance of these five times' tests is taken as the performance of 5-fold cross validation. A higher value of accuracy (or a lower value of Hamming loss) means that the feature selection method is more efficient.

B. Result Analysis

A summary of the nine public multi-label datasets is presented in the following Table I. As shown in Table I, the nine multi-label datasets come from many different fields, such as image, text, and biology, etc. These datasets' label counts range from 6 to 374. There are from 593 to 7395 observations and from 72 to 1836 features are present. We use these multi-label datasets to test the performance of the MAPD method.

TABLE I. Characteristics of the Nine Multi-Label Datasets

Name	Abbr.	*n*	*m*	*l*	LCard	LDen	Field
scene	Sce	2407	294	6	1.074	0.179	image
emotions	Emo	593	72	6	1.869	0.311	music
yeast	Yea	2417	103	14	4.237	0.303	biology
birds	Bir	645	260	19	1.014	0.053	audio
genbase	Gen	662	1186	27	1.252	0.046	biology
medical	Med	978	1449	45	1.245	0.028	text
enron	Enr	1702	1001	53	3.378	0.064	text
bibtex	Bib	7395	1836	159	2.402	0.015	text
corel5k	Cor	5000	499	374	3.522	0.009	image

Note: LCard means label cardinality; LDen means label density.

Label density is a standardized way to calculate label cardinality by dividing the total number of labels, where label cardinality is the average number of labels marked for each observation. The following formulas can be used to calculate label cardinality and label density: (Kashef and Nezamabadi-Pour, 2019):

$$LCard = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta(y_{ij} = 1) \quad (6)$$

$$LDen = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{m}\delta(y_{ij} = 1)}{l} \quad (7)$$

In the MAPD algorithm, we first analyze approximate the Pareto dominance of different approximate parameter α. Considering the range of the approximate parameter is $0.5<\alpha<1$, we set four groups of tests, i.e., α=0.6, α=0.7, α=0.8, and α=0.9. The results of Hamming loss for different approximate parameters are listed in Table II.

As shown in Table II, when the value of the approximate parameter decreases, the average of Hamming loss for the nine datasets decreases. Specifically, successively subtracting two adjacent items in the last row of Table II, we can see that the reduction of Hamming loss is 0.36%, 0.22%, and 0.07% for each 0.1 reduction of the approximate parameter α,

respectively. Therefore, the proposed MAPD algorithm has a trend of continuous reduction and convergence in the Hamming loss criterion. Specifically, the average of the Hamming loss is optional when the approximate parameter α=0.6. The optimal Hamming loss is obtained on 7, 5, and 3 of 9 datasets when the approximate parameters are α=0.6, α=0.7, and α=0.8, respectively. The standard deviations of Hamming loss are stable for all approximate parameters.

TABLE II. HAMMING LOSS FOR DIFFERENT APPROXIMATE PARAMETERS

Datasets	α=0.6	α=0.7	α=0.8	α=0.9
Sce	**0.0004** (0.00004)	**0.0004** (0.00005)	**0.0004** (0.00005)	0.0035 (0.00026)
Emo	**0.0057** (0.00073)	0.0069 (0.00068)	0.0069 (0.00068)	0.0236 (0.00037)
Yea	0.0068 (0.00026)	**0.0067** (0.00039)	**0.0067** (0.00029)	0.0068 (0.00016)
Bir	0.0534 (0.00000)	**0.0513** (0.00028)	0.0540 (0.00009)	0.0543 (0.00051)
Gen	**0.0028** (0.00023)	**0.0028** (0.00023)	**0.0028** (0.00025)	0.0030 (0.00013)
Med	**0.0032** (0.00013)	0.0036 (0.00010)	0.0063 (0.00013)	0.0084 (0.00021)
Enr	**0.0128** (0.00008)	0.0196 (0.00015)	0.0325 (0.00018)	0.0379 (0.00011)
Bib	**0.0043** (0.00001)	**0.0043** (0.00001)	0.0051 (0.00001)	0.0085 (0.00003)
Cor	**0.0027** (0.00001)	0.0028 (0.00001)	0.0035 (0.00001)	0.0045 (0.00002)
Mean	**0.0102** (0.00017)	0.0109 (0.00021)	0.0131 (0.00019)	0.0167 (0.00020)

Note: The number outside the bracket is the mean value of Hamming loss of 5-fold cross verifications of 5 times, and the number inside the bracket is standard deviations.

TABLE III. INFLUENCE OF DIFFERENT APPROXIMATE PARAMETERS ON THE ACCURACY

Datasets	α=0.6	α=0.7	α=0.8	α=0.9
Sce	**0.9977** (0.00023)	0.9975 (0.00029)	0.9975 (0.00029)	0.9795 (0.00127)
Emo	**0.9659** (0.00436)	0.9585 (0.00406)	0.9585 (0.00406)	0.8641 (0.00151)
Yea	0.9251 (0.00304)	**0.9259** (0.00295)	0.9244 (0.00218)	0.9231 (0.00229)
Bir	0.4558 (0.00000)	**0.4998** (0.00460)	0.4667 (0.00347)	0.4645 (0.00544)
Gen	0.9432 (0.00313)	0.9432 (0.00313)	**0.9447** (0.00274)	0.9390 (0.00135)
Med	**0.8613** (0.00466)	0.8521 (0.00406)	0.7513 (0.00418)	0.6847 (0.00757)
Enr	**0.5524** (0.00553)	0.4403 (0.00560)	0.2496 (0.00384)	0.2018 (0.00375)
Bib	**0.6127** (0.00275)	**0.6127** (0.00275)	0.5567 (0.00214)	0.3128 (0.00271)
Cor	**0.3089** (0.00265)	0.2952 (0.00246)	0.2430 (0.00113)	0.1664 (0.00283)
Mean	**0.7359** (0.00293)	0.7250 (0.00332)	0.6769 (0.00267)	0.6151 (0.00319)

The results of accuracy for different approximate parameters are shown in Table III. When the values of the approximate parameter reduce, the average value of the accuracy for the nine datasets increases. Similar to the calculation in Table II, for each 0.1 reduction of approximate parameters, the accuracy will increase by 6.18%, 4.81%, and 1.09%, respectively. The accuracy criterion of the proposed MAPD algorithm has an increasing and converging trend. Specifically, the mean value of the accuracy is optimal for the approximate parameter α=0.6. When the approximate parameters are α=0.6, α=0.7, and α=0.8, 6, 3, and 1 of 9 datasets get the optimal accuracy, respectively. The standard deviations of the accuracy criterion are stable.

Fig. 1 displays the outcome of the number of features chosen for various approximation values. We use the natural based logarithm of the size of the selected features when plotting the histogram. As the number of the selected features calculated by the proposed algorithm is 1 for several datasets (e.g., Gen, Cor and Yea), the scale of the chosen features in the Fig. 1 is zero. As shown in Fig. 1, when approximate parameter decreases, the number of chosen features by the proposed multi-label feature selection algorithm decreases. This finding is consistent with the definition of the proposed approximate Pareto dominance. When the value of the approximate parameter decreases, the number of dimensions that need to be satisfied is reduced for determining if one solution is better than the other or not. This makes it relatively easy to satisfy the concept of approximate Pareto dominance between solutions in the original set of solutions, decreasing the number of approximate Pareto dominated solutions and, in turn, the number of the chosen features of the suggested MAPD technique.

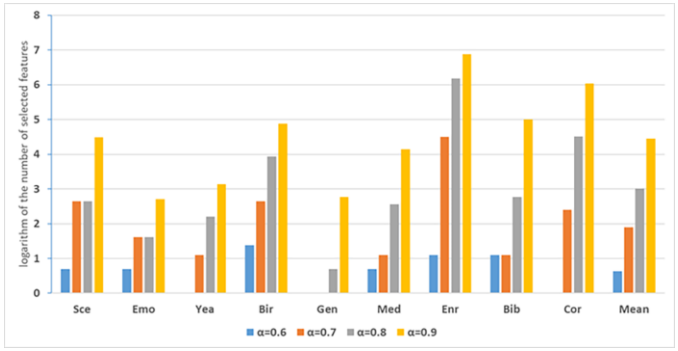

Fig. 1. Influence of different approximate parameters on the scale of the chosen features.

We also compare Hamming loss of the proposed MAPD method with other feature selection methods, i.e., ParFS, SCLS, and AMI. As shown in Table IV, the average value of Hamming loss of the proposed algorithm (MAPD) is optimal (0.0102). Compared with the Hamming loss values of ParFS, SCLS, and AMI algorithms, the Hamming loss of the MAPD method is reduced by 0.68%, 0.56%, and 0.18%, respectively. Specifically, the proposed MAPD algorithm obtains the optimal Hamming loss for the six multi-label datasets (i.e., Sce, Emo, Med, Enr, Bib, and Cor). However, other feature selection methods have a low performance of Hamming loss. Specifically, ParFS algorithm obtains the optimal Hamming loss for Yea dataset; SCLS algorithm obtains the optimal Hamming loss for Gen dataset; AMI algorithm obtains the optimal Hamming loss for Bir dataset. Moreover, for Cor, Bib, Enr and Med datasets with the highest number of labels, the proposed MAPD algorithm obtains the optimal Hamming loss.

Thus, it follows that datasets with high-dimensional labels are more suited for the suggested MAPD method. Compared with ParFS, SCLS, and AMI algorithms, the reduction ranges of Hamming loss of the proposed MAPD method are -0.02% ~ 20.99%, -0.15% ~ 2.36%, and -0.67% ~ 1.42%, respectively. The standard deviation values of the MAPD algorithm are similar to the other algorithms.

TABLE IV. HAMMING LOSS PERFORMANCE OF DIFFERENT METHODS

Datasets	ParFS	SCLS	AMI	MAPD
Sce	0.0035 (0.00026)	0.0010 (0.00015)	0.0011 (0.00008)	**0.0004** (0.00004)
Emo	0.0236 (0.00037)	0.0293 (0.00131)	0.0090 (0.00063)	**0.0057** (0.00073)
Yea	**0.0066** (0.00013)	0.0071 (0.00026)	0.0070 (0.00006)	0.0068 (0.00026)
Bir	0.0539 (0.00059)	0.0519 (0.00060)	**0.0467** (0.00058)	0.0534 (0.00000)
Gen	0.0030 (0.00011)	**0.0027** (0.00006)	0.0030 (0.00017)	0.0028 (0.00023)
Med	0.0087 (0.00017)	0.0066 (0.00025)	0.0046 (0.00017)	**0.0032** (0.00013)
Enr	0.0379 (0.00020)	0.0304 (0.00024)	0.0270 (0.00004)	**0.0128** (0.00008)
Bib	0.0107 (0.00008)	0.0100 (0.00002)	0.0070 (0.00001)	**0.0043** (0.00001)
Cor	0.0046 (0.00002)	0.0031 (0.00003)	0.0030 (0.00003)	**0.0027** (0.00001)
Mean	0.0170 (0.00022)	0.0158 (0.00033)	0.0120 (0.00020)	**0.0102** (0.00017)

TABLE V. PERFORMANCES OF DIFFERENT FEATURE SELECTION METHODS ON ACCURACY

Datasets	ParFS	SCLS	AMI	MAPD
Sce	0.9795 (0.00127)	0.9938 (0.00088)	0.9937 (0.00035)	**0.9977** (0.00023)
Emo	0.8641 (0.00151)	0.8331 (0.00773)	0.9460 (0.00377)	**0.9659** (0.00436)
Yea	**0.9258** (0.00162)	0.9214 (0.00265)	0.9236 (0.00179)	0.9251 (0.00304)
Bir	0.4667 (0.00190)	0.4695 (0.00318)	**0.5029** (0.00419)	0.4558 (0.00000)
Gen	0.9405 (0.00083)	**0.9465** (0.00274)	0.9375 (0.00172)	0.9432 (0.00313)
Med	0.6816 (0.00661)	0.7466 (0.00793)	0.8186 (0.00553)	**0.8613** (0.00466)
Enr	0.2001 (0.00287)	0.2599 (0.00552)	0.2979 (0.00208)	**0.5524** (0.00553)
Bib	0.2006 (0.00195)	0.2216 (0.00127)	0.4077 (0.00321)	**0.6127** (0.00275)
Cor	0.1596 (0.00291)	0.2666 (0.00455)	0.2730 (0.00529)	**0.3089** (0.00265)
Mean	0.6020 (0.00238)	0.6288 (0.00405)	0.6779 (0.00310)	**0.7359** (0.00293)

We compare the accuracy of different multi-label feature selection methods (see Table V). As shown in Table V, the proposed MAPD algorithm obtains the optimal average value of accuracy of nine datasets (0.7359). Compared with ParFS, SCLS, and AMI algorithms, the accuracy value of the proposed MAPD algorithm has increased by 13.39%, 10.71%, and 5.80%, respectively. The proposed MAPD algorithm obtains the optimal accuracy value for six datasets, i.e., Sce, Emo, Med, Enr, Bib, and Cor. ParFS algorithm obtains the optimal accuracy value for Yea dataset; SCLS algorithm obtains the optimal accuracy value for Gen dataset; AMI algorithm obtains the optimal accuracy value for Bir dataset. Moreover, for Cor, Bib, Enr and Med datasets with the highest number of labels, the proposed MAPD algorithm obtains the optimal accuracy value. Compared the ParFS, SCLS, and AMI algorithms, the increase ranges of the MAPD algorithm on the nine datasets are -1.09% ~ 41.21%, -1.37% ~ 39.11%, and -4.71% ~ 25.45%, respectively. In terms of the standard deviation, all algorithms are similar.

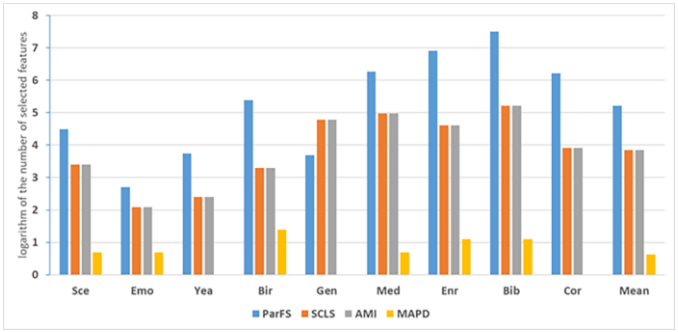

Fig. 2. The number of selected features of different methods.

In Fig. 2, the scale of the chosen features of the MAPD algorithm is much lower than the scale of the chosen features of the other three methods. When the dimensions of the labels increase, the advantage of the MAPD algorithm is strengthened.

V. CONCLUSIONS

For the high-dimensional feature selection problem with multi-label dataset, a concept called approximate Pareto dominance is presented, which can be used to compare the qualities of two solutions. Compared with the traditional Pareto dominance concept, with the help of the approximate parameter, the proposed approximate Pareto dominance can solve the problem that Pareto dominance cannot do well when the evaluation function dimension of the solution is high. Then, using this concept, the feature selection problem with multi-label data is mapped to the problem of finding the set of approximate Pareto dominance solutions. Based on this transformation, we propose a method called MAPD to solve it.

The MAPD algorithm is tested on nine public multi-label datasets from different fields. Experiment results show that the proposed MAPD algorithm has a higher level of classification accuracy, a lower level of Hamming loss and a lower number of the selected features compared with the existing methods. Specifically, compared with ParFS, SCLS and AMI methods, the Hamming loss evaluation index is reduced by 0.68%, 0.56% and 0.18%, respectively. The proposed MAPD method obtains the optimal Hamming loss on six of nine datasets. On the accuracy evaluation index, the proposed MAPD algorithm also obtains the best classification accuracy on six of nine datasets; compared with ParFS, SCLS and AMI methods, the classification accuracy increased by 13.39%, 10.71% and 5.80%, respectively.

REFERENCES

[1] Zhang, Y., Gong, D., Hu, Y., & Zhang, W. (2015). Feature selection algorithm based on bare bones particle swarm optimization. Neurocomputing, 148, 150-157.

[2] Pashaei, E., & Aydin, N.. (2017). Binary black hole algorithm for feature selection and classification on biological data. Applied Soft Computing, 56, 94-106.

[3] Too, J., & Mirjalili, S.. (2020). General learning equilibrium optimizer: a new feature selection method for biological data classification. Applied Artificial Intelligence, 35(3), 1-17.

[4] Qiao, L., Zhang, L., Sun, Z., & Liu, X.. (2017). Selecting label-dependent features for multi-label classification. Neurocomputing, 259, 112-118.

[5] Siblini, W., Kuntz, P., & Meyer, F.. (2021). A review on dimensionality reduction for multi-label classification. IEEE Transactions on Knowledge and Data Engineering, 33(3), 839-857.

[6] Yang, J., Jiang, Y. G., Hauptmann, A. G., & Ngo, C. W. (2007). Evaluating bag-of-visual-words representations in scene classification. Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval, ACM, MIR 2007, Augsburg, Bavaria, Germany, September 24-29.

[7] Cabral, R., De, l. T. F., Costeira, J. P., & Bernardino, A. (2015). Matrix completion for weakly-supervised multi-label image classification. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37(1), 121-35.

[8] Jiang, J. Y., Tsai, S. C., & Lee, S. J. (2012). Fsknn: multi-label text categorization based on fuzzy similarity and k nearest neighbors. Expert Systems with Applications, 39(3), 2813-2821.

[9] Elghazel, H., Aussem, A., Gharroudi, O., & Saadaoui, W. (2016). Ensemble multi-label text categorization based on rotation forest and latent semantic indexing. Expert Systems with Applications, 57(Sep.), 1-11.

[10] Wu, J. S., Huang, S. J., & Zhou, Z. H. (2014). Genome-wide protein function prediction through multi-observation multi-label learning. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 11(5), 891-902.

[11] Lee, J., & Kim, D. W. (2013). Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters, 34(3), 349-357.

[12] Reyes, O. G., Morell, C., & Ventura, S. (2015). Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context. Neurocomputing, 161(aug.5), 168-182.

[13] Li, F., Miao, D. Q., & Pedrycz, W. (2017). Granular multi-label feature selection based on mutual information. Pattern Recognition, 67, 410-423.

[14] Yu, Y., & Wang, Y. L. (2014). Feature Selection for Multi-label Learning Using Mutual Information and GA. International Conference on Rough Sets and Knowledge Technology. Springer International Publishing, 454-463.

[15] Lee, J., & Kim, D. W. (2015). Memetic feature selection algorithm for multi-label classification. Information Sciences, 293, 80-96.

[16] You, M. Y., Liu, J. M., Li, G. Z., Chen, Y. (2012). Embedded Feature Selection for Multi-label Classification of Music Emotions. International Journal of Computational Intelligence Systems, 5(4), 668-678.

[17] Zhu, P. F., Xu, Q., Hu, Q. H., Zhang, C. Q., & Zhao, H. (2016). Robust Multi-label Feature Selection with Missing Labels. Chinese Conference on Pattern Recognition. Springer, Singapore, 662, 752-765.

[18] Doquire, G., & Verleysen, M. Feature selection for multi-label classification problems (2011). International Work-Conference on Artificial Neural Networks, pp. 9-16.

[19] Spolaor, N., Cherman, E. A., Monard, M. C., & Lee, D. L. (2013). A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach. Electronic Notes in Theoretical Computer Science, 292, 135-151.

[20] Doquire, G., & Verleysen, M. (2013). Mutual information-based feature selection for multi-label classification. Neurocomputing, 122,148-155.

[21] Lin, Y. J., Hu, Q. H., Liu, J. H., & Duan, J. (2015). Multi-label feature selection based on max-dependency and min-redundancy. Neurocomputing, 168, 92-103.

[22] Zhang, M. L., Pena, J. M., & Robles, V. (2009). Feature selection for multi-label naive Bayes classification. Information Sciences, 179(19), 3218-3229.

[23] Kong, X. N., & Yu, P. S. (2012). gMLC: a multi-label feature selection framework for graph classification. Knowledge & Information Systems, 31(2), 281-305.

[24] Li, P., Li, H., & Wu, M. (2013). Multi-label ensemble based on variable pairwise constraint projection. Information Sciences, 222(3), 269-281.

[25] Pupo, O. G. R., Morell, C., & Soto, S. V. (2013). ReliefF-ML: An Extension of ReliefF Algorithm to Multi-label Learning. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications.

[26] Lee, J., Lim, H., & Kim, D. W. Approximating mutual information for multi-label feature selection. Electronics Letters, 2012, 48(15), 929-930.

[27] Lee, J., & Kim, D. W. SCLS: Multi-label feature selection based on scalable criterion for large label set. Pattern Recognition, 2017, (66), 342–352.

[28] Kashef, S., & Nezamabadi-Pour, H. (2015). An advanced ACO algorithm for feature subset selection. Neurocomputing, 147, 271-279.

[29] Kashef, S., & Nezamabadi-Pour, H. (2019). A label-specific multi-label feature selection algorithm based on the Pareto dominance concept. Pattern Recognition, 88, 654-667.

[30] Fan, Y. L., Chen, B. H., Huang, W. Q., Liu, J. H., Weng, W., & Lan, W. Y. (2022). Multi-label feature selection based on label correlations and feature redundancy. Knowledge-based systems, 241, 108256.

[31] Hu, L., Gao, L. B., Li, Y. H., Zhang, P., & Gao, W. F. (2022). Feature-specific mutual information variation for multi-label feature selection. Information Sciences, 593, 449-471.

[32] Kannan, S. S., & Ramaraj, N. (2010). A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm. Knowledge-Based Systems, 23(6), 580-585.

[33] Sosa-Cabrera, G., Garcia-Torres, M., Gomez-Guerrero, S., Schaerer, C. E., & Divina, F. (2019). A multivariate approach to the symmetrical uncertainty measure: application to feature selection problem. Information Sciences, 494.

[34] Xue, Y. N., Li, M. Q., Shepperd, M., Lauria, S., & Liu, X. H.. (2019). A novel aggregation-based dominance for pareto-based evolutionary algorithms to configure software product lines. Neurocomputing, 364, 32-48.

[35] Boutell, M. R. , Luo, J. , Shen, X. , & Brown, C. M. (2004). Learning multi-label scene classification. Pattern Recognition, 37(9), 1757-1771.

[36] Trohidis, K. , Tsoumakas, G. , Kalliris, G. , & Vlahavas, I. P. (2008). Multi-label Classification of Music into Emotions. International Conference on Music Information Retrieval (ISMIR 2008), pp. 325-330, Philadelphia, PA, USA.

[37] Briggs, F. , Huang, Y. , Raich, R. , Eftaxias, K. , & Milakov, M. (2013). The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment. IEEE International Workshop on Machine Learning for Signal Processing. IEEE.

[38] Cherman, E. A., Spolaor, N., Valverde-Rebaza, J., & Monard, M. C. (2015). Lazy multi-label learning algorithms based on mutuality strategies. Journal of Intelligent & Robotic Systems, 80(1), 261-276.

[39] Elkafrawy, P., Mausad, A., & Esmail, H. (2015). Experimental comparison of methods for multi-label classification in different application domains. International Journal of Computer Applications, 114(19), 1-9.