Multi-label Feature Selection Method via Maximizing Correlation-based Criterion with Mutation Binary Bat Algorithm

Yuanyuan Tao
School of Computer Science and Technology, Nanjing Normal University
Nanjing, Jiangsu, China
39905635@qq.com

Jun Li
School of Computer Science and Technology, Nanjing Normal University
Nanjing, Jiangsu, China
lijuncst@njnu.edu.cn

Jianhua Xu*
School of Computer Science and Technology, Nanjing Normal University
Nanjing, Jiangsu, China
xujianhua@njnu.edu.cn

***Abstract*—Multi-label feature selection is a vital pre-processing step to reduce computational complexity, improve classification performance and enhance model interpretability, via selecting a discriminative subset of features from original high-dimensional features. Correlation-based feature selection (CFS) criterion measures the relevance between features and labels, and the redundancy among features, which has been combined with hill climbing and genetic algorithm to execute multi-label feature selection task. However, it is an open problem to search for more effective optimization tools for CFS. In this paper, through adding a mutation operation, we modify existing binary bat algorithm to build its mutation version (MBBA), to adjust the number of "1" components to be a fix size. Then a new multi-label feature selection approach is proposed via maximizing CFS criterion using MBBA, to select a fixed number of discriminative features. Our experiments on four data sets show that our proposed method is superior to three state-of-the-art approaches, according to four sample-based performance evaluation metrics for multi-label classification.**

***Index Terms*—multi-label learning, feature selection, correlation-based criterion, bat algorithm, mutation operation**

I. Introduction

The pattern classification is to build a classification model using labeled training samples and then to predict class labels for unannotated samples [1], [2]. According to the number of labels corresponding to a single sample, classification problems can be divided into two categories: single label and multi-label ones. Each sample is related to only one label in the former case, and multiple labels at the same time in the latter one [3]. Nowadays, there are many multi-label application areas, for example, text categorization, music emotion classification, and image annotation [3]. In Fig. 1, ten images from [4] are shown, where the first row indicates five single-label images (desert, mountain, sea, sunset and tree) and the second row lists five multi-label images annotated by the above five labels.

With the development of various sensor and related post-processing techniques, a large number of features are available, which unavoidably include some redundant and irrelevant features in multi-label classification [5]. Such a high-dimensional data would increase computational complexity, and even degrade classification performance, which could be dealt with using feature selection (FS) strategy to select a small subset of discriminative features from original high-dimensional features [5], [6]. Compared with the single-label FS task, multi-label FS one is more difficult since there exist more complicated relevances between features and labels, and correlations among labels [7]. Therefore, recently multi-label FS task becomes a hot issue in pattern recognition, machine learning, big data and so on [5], [6], [8].

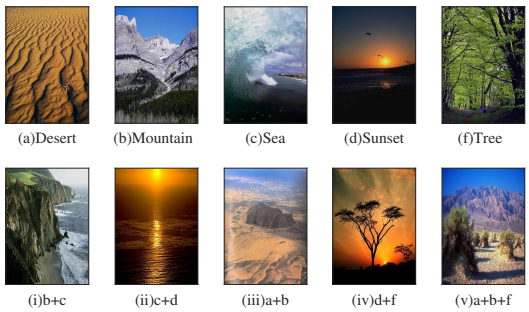

Fig. 1. Ten original images from Image data set

According to the intersection between FS and classifier in the implementation process, existing FS methods can generally be divided into three categories: filter, wrapper and embedded [9]. Filter methods evaluate the quality of features on the basis of the intrinsic characteristics and structures of data, without using any learning algorithm [7], [10]. Wrapper techniques need a proper classifier to estimate the classification performance to evaluate the quality of selected features [11], [12]. Embedded methods integrate feature selection into classifier design and implementation [13], [14]. Generally, the first kind of methods are more efficient than the last two kinds of techniques, which results in that more attention has been pain

This work was supported by the Natural Science Foundation of China (NSFC) under Grants 61273246 and 61703096.
*Corresponding author

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

to filter-based approaches in multi-label classification.

A filter-based FS method consists of two parts: a proper evaluation index and a related optimization technique. There are mainly three kinds of optimization strategies: simple ranking, greedy search and swarm intelligence. F-statistic and ReliefF criteria [10], and mutual information-based criterion [15], are used to evaluate the quality of each feature, and then some top ranked features are selected through simple ranking way. The widely used greedy search methods include hill climbing and sequential forward selection. In [16], correlation-based criterion is combined with hill climbing. Sequential forward selection is to optimize Hilbert-Schmidt independence criterion (HSIC) variant in [17] and mutual information type indexes in [18], [19], to create the sub-optimal subset of features. To search for a globally optimal feature subset, genetic algorithm is used to maximize HSIC [20] and correlation-based criterion [21], while particle swarm optimization optimizes mutual information index [22]. It is still an open problem to find out a more effective swarm intelligence algorithm for a proper criterion in multi-label FS field.

Correlation-based feature selection (CFS) criterion originally for single-label FS [23] is to maximize relevance between features and labels, and minimize redundancy among features, which has been extended to multi-label feature selection via hill climbing [16] and genetic algorithm [21]. It is experimentally observed that the general genetic algorithm prefers to select a half of original features [12]. Bat algorithm (BA) is one of swarm intelligence algorithms for continuous variables, which simulates bat echolocation behavior with global and local search strategies [24]. To fit binary variables, its binary version (BBA) is generalized in [25], which also could not control the number of "1" components. In this paper, we add a mutation operation used in genetic algorithm originally, to construct a new BBA version (i.e., MBBA). Then MBBA is applied to maximize CFS criterion to propose a novel multi-label FS algorithm (simply CFS-BA), to select a fixed size subset of discriminative features. Finally, experiments on four benchmark data sets show that our proposed method works better, compared with three existing methods in [15], [21], [22], according to four sample-based evaluation metrics for multi-label classification [3].

The rest of this paper is organized as follows. In Section 2, we introduce a novel multi-label FS method. Section 3 is associated with our experiments and analysis. Finally, we conclude our research work in Section 4.

II. Novel Multi-label Feature Selection Method

In this section, a new multi-label FS is proposed, which is described using four parts: preliminaries, correlation-based FS criterion, mutation binary bat algorithm, and novel multi-label FS method.

A. Preliminaries

Assume that a given training data set is denoted by its real feature matrix $\mathbf{X} \in \mathcal{R}^{D\times N}$ and binary label matrix $\mathbf{Y} \in \{0,1\}^{L\times N}$ as follows

$$\begin{aligned}\mathbf{X} &= [\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_N] = \left[\mathbf{x}^1, ..., \mathbf{x}^j, ..., \mathbf{x}^D\right]^T \\ \mathbf{Y} &= [\mathbf{y}_1, ..., \mathbf{y}_i, ..., \mathbf{y}_N] = \left[\mathbf{y}^1, ..., \mathbf{y}^j, ..., \mathbf{y}^L\right]^T\end{aligned} \tag{1}$$

where *N*, *D* and *L* represent the numbers of samples, features and labels, respectively. The i-th sample is described using its column feature vector $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{iD}]^T \in \mathcal{R}^D$ and label vector $\mathbf{y}_i = [y_{i1}, y_{i2}, ..., y_{iL}]^T \in \{0,1\}^L$ ($y_{ij} = 1$ means that the j-th label is relevant). Moreover, $\mathbf{x}^j = [x_{1j}, x_{2j}, ..., x_{Nj}]^T$ and $\mathbf{y}^j = [y_{1j}, y_{2j}, ..., y_{Nj}]^T$ indicate the *j*-th feature and label vectors, respectively. The original feature index set is set to $\mathcal{F} = \{1, 2, ..., D\}$.

The multi-label feature selection is to choose a feature subset $\mathcal{S}$ of size *d* from the original $\mathcal{F}$ ($d < D$) to remain those highly relevant and lowly redundant features. To achieve this task, we adopt correlation-based feature selection (CFS) [23] criterion to search for a globally optimal subset via binary bat algorithm (BBA) [25] with mutation operator, in this paper.

B. Correlation-based Multi-label Feature Selection Criterion

Correlation-based feature selection criterion (CFS) is firstly proposed in [23] for single-label learning, which not only maximizes the correlations between features and labels, but also minimizes the redundances among features. Let $C_{fl}(\mathbf{x}^i, \mathbf{y}^j)$ be the correlation measure between *i*-th feature and *j*-th label, and $R_{ff}(\mathbf{x}^i, \mathbf{x}^j)$ be the redundance measure between *i*-th feature and *j*-th one, which could be calculated by Pearson correlation, mutual information, symmetrical uncertainty and so on. For a selected feature subset $\mathcal{S}$ from $\mathcal{F}$, the average correlation between the *i*-the feature and the entire label set is defined as

$$\bar{C}_{fL}(\mathbf{x}^i, \mathbf{Y}) = \frac{1}{L}\sum_{k=1}^{L} C_{fl}(\mathbf{x}^i, \mathbf{y}^k) \tag{2}$$

and the overall average correlation between those selected features and all labels as

$$\bar{C}_{SL} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \bar{C}_{fL}(\mathbf{x}^i, \mathbf{Y}) = \frac{1}{L\,|\mathcal{S}|}\sum_{i\in\mathcal{S}}\sum_{k=1}^{L} C_{fl}(\mathbf{x}^i, \mathbf{y}^k) \tag{3}$$

where $|\cdot|$ indicates the size of $\mathcal{S}$. The overall redundancy ($\bar{R}_{SS}$) among selected features are averaged across all possible feature pairs, i.e.,

$$\bar{R}_{SS} = \frac{1}{|\mathcal{S}|(|\mathcal{S}|-1)}\sum_{i,j\in\mathcal{S}, i\neq j} R_{ff}(\mathbf{x}^i, \mathbf{x}^j). \tag{4}$$

In this case, the CFS criterion [23] for single-label feature selection is defined as

$$\mathrm{CFS}(\mathcal{S}) = \frac{|\mathcal{S}|\bar{C}_{SL}}{\sqrt{|\mathcal{S}| + |\mathcal{S}|(|\mathcal{S}|-1)\,\bar{R}_{SS}}}. \tag{5}$$

In [23], the best first search method is applied for searching for the *d* features. The above criterion is also extended to multi-label case, in which hill climbing search in [16] and genetic algorithm (GA) in [21] are applied, respectively. The GA could find out a better subset than heuristic approaches.

Algorithm 1 A multi-label feature selection method via combining CFS with MBBA

Input

$\mathbf{X}$ and $\mathbf{Y}$: feature data and label matrices.
M: the size of bat population.
T: the maximal number of iterations.
d: the number of selected features.

Procedure

To initialize five quantities for each bat: $\mathbf{s}_i(0)$, $\mathbf{v}_i(0)$, $f_i(0)$, $r_i(0)$, and $a_i(0)$ $(i = 1, ..., M)$.
To set $t = 1$.
Repeat
For *i*=1 to *M* do
To generate a global new solution $\mathbf{s}_i(t)$ according to (8).
If a current random number $r \sim U(0,1) > r_i$ then to create a local new solution $\mathbf{s}_i(t)$ via (9).
If $r \sim U(0,1) < a_i$ and $g(\mathbf{s}_i(t)) > g(\mathbf{s}^*)$ then
to accept the new solution $\mathbf{s}_i(t)$, and to update $r_i(t+1)$ and $a_i(t+1)$ via (10) and (11).
To adjust the number of selected features to be *d* using mutation operator.
t=t+1.
Until $(t > T)$
To detect an optimal subset S of selected features using the highest $g(\mathbf{s})$.

Output:

S: the subset of selected features.

However, the GA could not give a fixed number of selected features in practice. In this study, we apply binary bat algorithm [25] to execute feature selection and moreover add a mutation operation to control the number of selected features.

C. Binary Bat Algorithm with Mutation Operator

Bat algorithm (BA) is a swarm intelligence algorithm inspired by bat echolocation behavior [24], [26], [27]. This algorithm initializes a set of random solutions and then optimizes them iteratively. Specially, some local optimal solutions are added through random flight around the optimal solution, to strengthen the local search ability. Compared with other algorithms (for example, genetic algorithm and particle swarm optimization), BA is much better in accuracy and effectiveness, and has no many parameters to be adjusted [24].

Assume that: (a) the size of bat population is set to *M*; (b) the dimensions of solution is *D*; (c) the position of the i-th bat is denoted by $\mathbf{s}_i = [s_{i1},..., s_{ij}, ..., s_{iD}]^T$ (*i*=1,..., *M*); (d) the moving velocity for the i-th bat is depicted as $\mathbf{v}_i = [v_{i1}, ..., v_{ij}, ..., \mathrm{v}_{iD}]^T$(*i*=1,...,*M*); and (e) the pulse frequency, pulse emission rate and loudness of the i-th bat are indicated f_i, a_i and r_i, respectively. Additionally, let $U(a, b)$ be a uniform distribution law in the real interval $[a, b]$.

The BA consists of two parts: global and local search. For the global search part, after randomly creating the initial position and velocity for the i-th bat, at the *t*-th time step, the original BA updates the following quantities:

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta \tag{6}$$

$$\mathbf{v}_i(t) = \mathbf{v}_i(t-1) + [\mathbf{s}^* - \mathbf{s}_i(t-1)]f_i \tag{7}$$

$$\mathbf{s}_i(t) = \mathbf{s}_i(t-1) + \mathbf{v}_i(t) \tag{8}$$

where f_{min} and f_{max} are the minimum and maximum frequencies, respectively, $\beta \sim U(0,1)$ denotes a random number satisfying $U(0,1)$, and $\mathbf{s}^* = [s_1^*, ..., s_j^*, ..., s_D^*]^T$ represents the current global optimal solution.

When a random number $r \sim U(0,1) > r_i$, a local search for the *i*-th bat is executed. After a solution ($\mathbf{x}^{old}$) is randomly selected from the current optimal solution set, its new solution ($\mathbf{x}^{new}$) is generated via the following random walk way:

$$\mathbf{s}^{new} = \mathbf{s}^{old} + \bar{a}(t)\mathbf{w} \tag{9}$$

where $\bar{a}(t)$ represents the average loudness of all the bats at this time step and $\mathbf{w} \sim U(-1,1)$ is a random vector which attempts to adjust the direction and strength of random walk. Finally let $\mathbf{s}_i(t) = \mathbf{s}^{new}$.

When a random number $r \sim U(0,1) < a_i$ and $g(\mathbf{s}_i(t)) > g(\mathbf{s}^*)$, we accept the solution $\mathbf{s}_i(t)$ from the above global or local search step, and then update the loudness a_i and emission pulse rate r_i as follows:

$$a_i(t+1) = \alpha a_i(t) \tag{10}$$

$$r_i(t+1) = r_i(0)[1 - exp(-\gamma t)] \tag{11}$$

where α and γ are two pre-defined positive constants, and the initial loudness $a_i(0) \sim U(1,2)$ and emission rate $r_i(0) \sim U(0,1)$.

For feature selection, the previous continuous solution $\mathbf{s}_i$ should be reduced into a binary one, whose "1" or "0" component represents that a feature is selected or not. Therefore binary bat algorithm (BBA) was proposed in [25], which uses the following sigmoid function:

$$\phi(v_{ij}) = \frac{1}{1 + e^{-v_{ij}}} \tag{12}$$

TABLE I
STATISTICAL INFORMATION FOR FOUR BENCHMARK MULTI-LABEL DATA SETS.

Dataset	#Domain	#Instances	#Features	#Classes	#Average labels
Scene	Image	2407	294	6	1.07
Image	Image	2000	294	5	1.24
Yeast	Biology	2417	103	14	4.24
Emotions	Music	593	72	6	1.87

TABLE II
THE NUMBER OF WINS FOR EACH METHOD AND METRIC ACROSS FOUR DATA SETS

Metric	CFS-GA	FIMF	MMI-PSO	CFS-BA
Accuracy	10	2	3	**61**
Subset Accuracy	13	2	2	**59**
Hamming loss	12	1	4	**59**
F1	12	0	0	**64**
Total wins	47	5	9	**243**

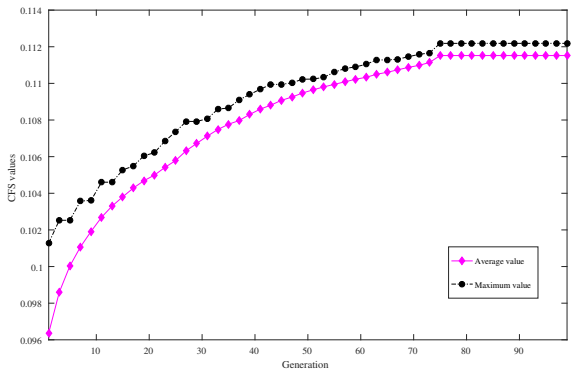

Fig. 2. Convergence analysis of CFS-BA on Image

to restrict each bat solution to binary form

$$s_{ij} = \begin{cases} 1 & \text{if } \phi(v_{ij}) > \sigma \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $\sigma \sim \mathrm{U}(0,1)$.

In this paper, to choose a fixed size for selected feature subset, we adjust the number of selected features to be d with mutation operator in genetic algorithm originally, which is also applied in [22]. Let the number of selected features be $\hat{d}$. When $d > \hat{d}$, the $(d - \hat{d})$ "0" components are selected randomly and then converted into "1" components. Reversely, we force the $(\hat{d} - d)$ "1" elements to be "0"s. In this study, the above BBA version with mutation operator is referred to as MBBA simply.

D. Multi-label feature selection method based on CFS and MBBA

In this subsection, we apply our MBBA to optimize CFS criterion (5) to build a new multi-label feature selection method (namely CFS-BA), as shown in **Algorithm 1**, where $g(\mathbf{s})$ represents the CFS criterion (5) and the selected feature subset $\mathcal{S}$ consists of those feature indexes with "1" components in $\mathbf{s}$. Finally, we select the best binary vector $\mathbf{s}^*$ with the largest $g(\mathbf{s}^*)$ to be our feature selection solution.

III. EXPERIMENTS

In this section, we evaluate our feature selection CFS-BA using four multi-label data sets via comparing with three existing methods.

A. Four Benchmark Data Sets

In this paper, we downloaded four widely-validated benchmark data sets: Scene, Image, Yeast and Emotions form [1], to evaluate and compare our algorithm and other existing feature selection methods, as shown in Table I. This table also shows some important statistics for these sets, including the numbers of samples, features, the size of labels, average labels, and application fields.

B. Compared Methods and Their Key Parameter Settings

In this study, we compare our CFS-BA with CFS-GA [21], FIMF [15] and MMI-PSO [22]. For three compared approaches, we accept their default settings. On our CFS-BA, its key parameters are assigned as follows: $f_{min} = 0$ and $f_{max} = 1$ recommended in [25], $\alpha = \gamma = 0.9$ used in [24], $T = 100$ and $M = 100$. The CFS criterion (5) is estimated via symmetrical uncertainty (SU) [23], whose two key components C_{fl} and R_{ff} are defined as follows:

$$C_{fl}(\mathbf{x}^i, \mathbf{y}^j) = 2\frac{H(\mathbf{x}^i) + H(\mathbf{y}^j) - H(\mathbf{x}^i, \mathbf{y}^j)}{H(\mathbf{x}^i) + H(\mathbf{y}^j)} \tag{14}$$

$$R_{ff}(\mathbf{x}^i, \mathbf{x}^j) = 2\frac{H(\mathbf{x}^i) + H(\mathbf{x}^j) - H(\mathbf{x}^i, \mathbf{x}^j)}{H(\mathbf{x}^i) + H(\mathbf{x}^j)} \tag{15}$$

where $H(\cdot)$ is the entropy measure in [23], [28], [29], and for some continuous feature vector $\mathbf{x}^i$, its components are binarized into 1 ($\geq \mu_i$) or 0 ($< \mu_i$) according to its corresponding mean value μ_i.

C. Baseline Classifier and Evaluation Metrics

In order to compare four aforementioned feature selection method, it is needed to select a proper baseline classifier and some multi-label classification performance indexes.

In this study, the widely-used multi-label k nearest neighbor method (ML-kNN) with a recommended k = 10 [4] is considered as our baseline classifier.

The multi-label classification performance is evaluated via four sample-based metrics, including accuracy, Hamming loss, subset accuracy and F1. For a test sample vector $\mathbf{x}$, its binary actual and predicted label vectors are denoted by $\mathbf{y} = [y_1, ..., y_j, ..., y_L]^T \in \{0,1\}^L$ and

[1] http://computer.njnu.edu.cn/Lab/LABIC/LABIC Software.html

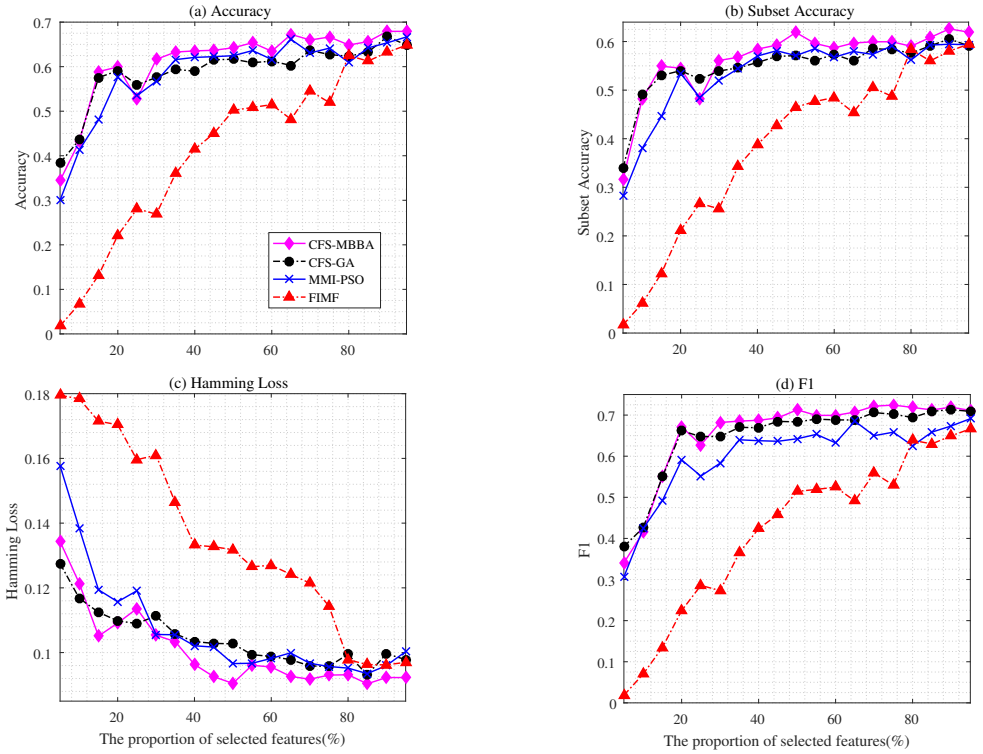

Fig. 3. Four sample-based metrics from four FS methods on Scene

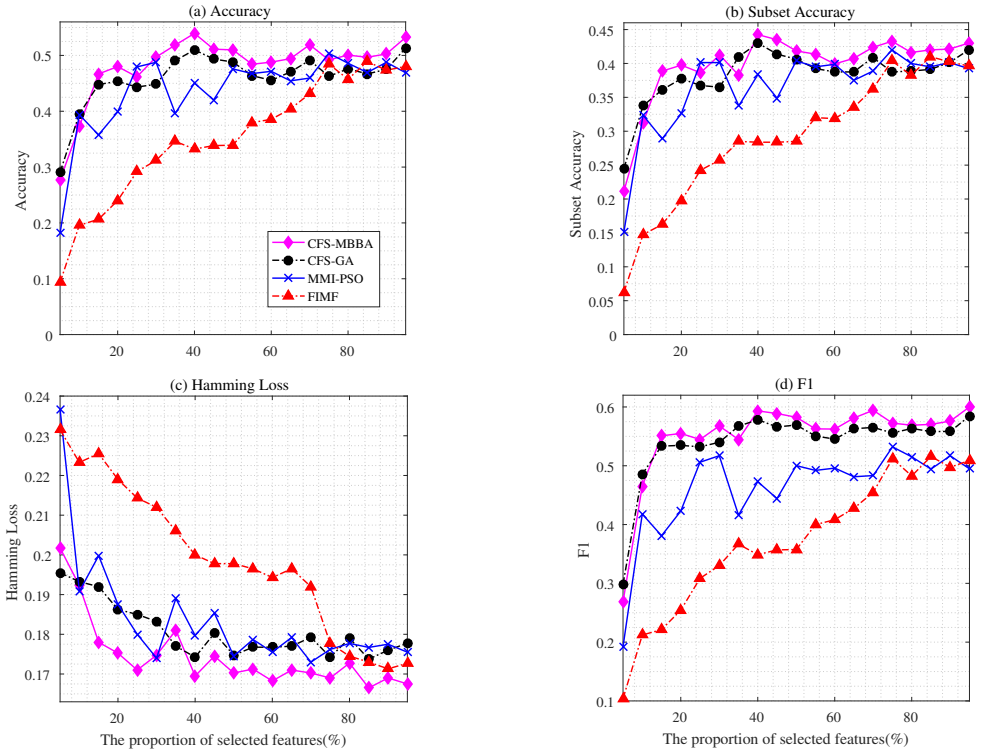

Fig. 4. Four sample-based metrics from four FS methods on Image

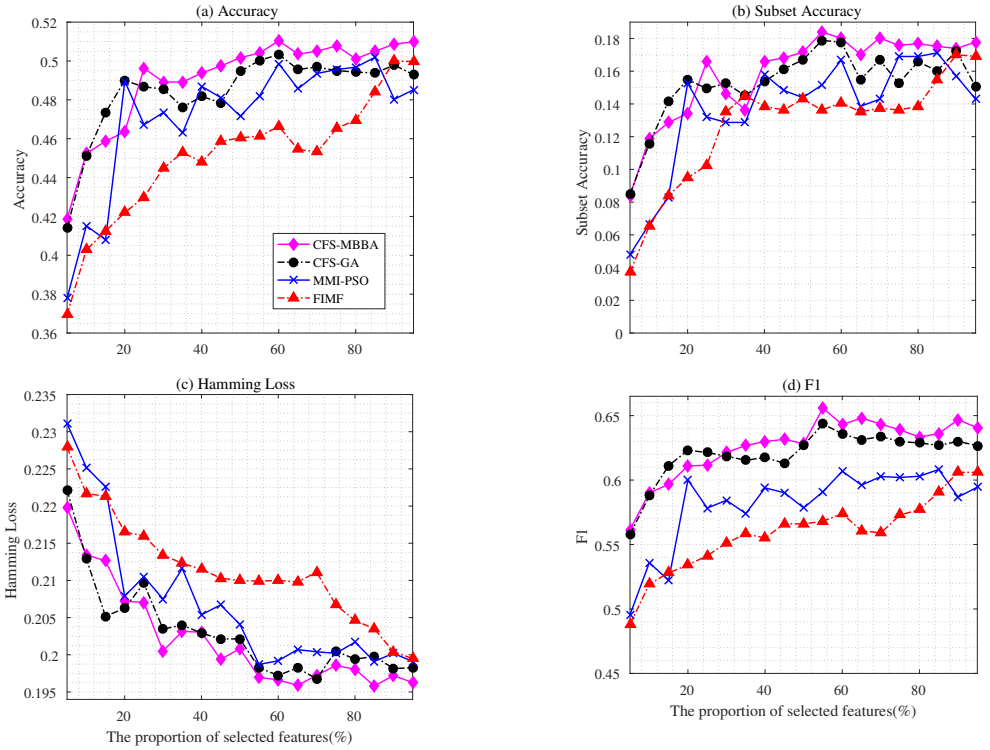

Fig. 5. Four sample-based metrics from four FS methods on Yeast

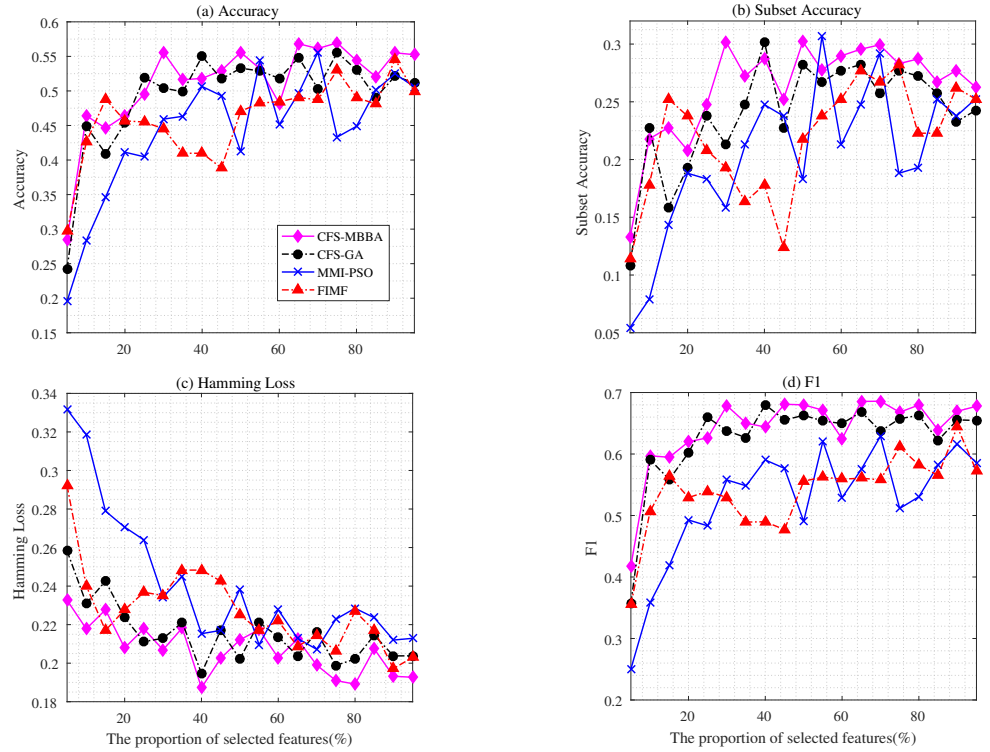

Fig. 6. Four sample-based metrics from four FS methods on Emotions

$\bar{\mathbf{y}} = [\bar{y}_1, ..., \bar{y}_j, ..., \bar{y}_L]^T \in \{0,1\}^L$, respectively. The first three metrics are defined as following forms:

TABLE III
FOUR METRICS FROM FOUR METHODS AND FOUR DATA SETS WITH 50% SELECTED FEATURES

Metric	CFS-GA	FIMF	MMI-PSO	CFS-BA
Scene				
Accuracy(↑)	0.616±0.001	0.501±0.001	0.624±0.002	**0.642±0.001**
Subset accuracy(↑)	0.571±0.002	0.464±0.001	0.572±0.001	**0.618±0.001**
Hamming loss(↓)	0.102±0.001	0.131±0.001	0.096±0.001	**0.090±0.002**
F1(↑)	0.683±0.003	0.514±0.002	0.642±0.002	**0.712±0.001**
Image				
Accuracy(↑)	0.488±0.002	0.339±0.001	0.475±0.001	**0.509±0.001**
Subset accuracy(↑)	0.406±0.001	0.285±0.002	0.403±0.001	**0.418±0.002**
Hamming loss(↓)	0.175±0.001	0.197±0.001	0.174±0.001	**0.170±0.001**
F1(↑)	0.569±0.001	0.357±0.001	0.501±0.002	**0.582±0.001**
Yeast				
Accuracy(↑)	0.494±0.001	0.460±0.001	0.471±0.001	**0.501±0.001**
Subset accuracy(↑)	0.166±0.002	0.142±0.002	0.143±0.001	**0.171±0.002**
Hamming loss(↓)	0.202±0.001	0.211±0.001	0.204±0.001	**0.201±0.001**
F1(↑	0.626±0.002	0.578±0.002	0.565±0.001	**0.628±0.002**
Emotions				
Accuracy(↑)	0.532±0.001	0.471±0.002	0.412±0.001	**0.556±0.001**
Subset accuracy(↑)	0.282±0.002	0.217±0.001	0.183±0.001	**0.302±0.003**
Hamming loss(↓)	**0.202±0.001**	0.225±0.002	0.238±0.002	0.212±0.001
F1(↑)	0.662±0.003	0.555±0.003	0.491±0.002	**0.681±0.002**

$$\text{Accuracy}(\uparrow) = \frac{\sum_{j=1}^{L} y_j \bar{y}_j}{\sum_{j=1}^{L} (y_j + \bar{y}_j - y_j \bar{y}_j)} \tag{16}$$

$$\text{Hamming loss}(\downarrow) = \frac{\sum_{j=1}^{L} (y_j - \bar{y}_j)^2}{L} \tag{17}$$

$$\text{Subset accuracy}(\uparrow) = \begin{cases} 1, & \text{if Hamming loss} = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

With two proxy metrics: precision and recall, the F1 metric is defined as

$$\text{Precision} = \frac{\sum_{j=1}^{L} y_j \bar{y}_j}{\sum_{j=1}^{L} \bar{y}_j} \tag{19}$$

$$\text{Recall} = \frac{\sum_{j=1}^{L} y_j \bar{y}_j}{\sum_{j=1}^{L} y_j} \tag{20}$$

$$\text{F1}(\uparrow) = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{21}$$

Except for Hamming loss, the higher the other metric values are, the better the feature selection methods work, as shown in upper and down arrows (↑ and ↓) in the above definitions.

For a testing set, the above four metrics are averaged across all testing samples. Additionally, since ten-fold cross validation is executed, the mean and standard deviation format is reported in our experiments.

D. Convergence Analysis for CFS-BA

In this sub-section, we regard the CFS criterion (5) as a function of the number of generations to investigate the convergence of our CFS-BA on Image, where the 50% features is selected (i.e., $|\mathcal{S}| = 147$), as shown in Fig. 2. As the number of generations increases from 1 to 100, the average and maximum CFS values ascend correspondingly and finally tend to be stable, which illustrates that our proposed method is convergent experimentally.

E. Experimental Results and Analysis

In this sub-section, we report our experimental results from four methods and four data sets. To evaluate the classification performance of each method comprehensively, we regard each metric as a function of the proposition of selected features from 5% to 95% with a step 5%, as shown in Figs. 3-6.

From these four figures, it is observed that at most of propositions of selected features, our CFS-BA performs the best. To compare these four FS methods in detail, we use "win" index [30] to count the number of the best results for each method and each metric across four data sets and 19 propositions (76 combinations), as shown in Table II. Among the total 304 wins, our CFS-BA achieves 243 ones, which is much greater than those from the other three methods.

To obtain a more extensive comparison, we list four metric values from four methods and four data sets with 50% selected features in Table III, where the best result is shown in bold font for each data and each metric. It is found out that our CFS-BA obtains 15 best metric values, and only CFS-GA performs the best on Hamming loss from Emotions.

Overall, the above experimental results and analysis demonstrate the effectiveness of our proposed method, compared with three state-of-the-art methods.

IV. CONCLUSIONS

In this paper, to control the number of selected features, we modify the traditional binary bat algorithm via adding a mutation operation, to present a new bat algorithm version for feature selection in particular. Combining such an optimization technique with correlation-based feature selection criterion is to construct a novel multi-label feature selection approach. Experiments from four data sets and four evaluation metrics (accuracy, Hamming loss, subset accuracy and F1) illustrate that our proposed method is superior to three existing methods.

For future work, we will compare our method with more existing approaches on more benchmark data sets according to more evaluation metrics.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley and Sons, 2001.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Singapore: Springer, 2006.

[3] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel Classification Problem Analysis, Metrics and Techniques*. Switzerland: Springer, 2016.

[4] M.-L. Zhang and Z.-H. Zhou, "ML-kNN: a lazy learning approach to multilabel learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.

[5] R. Pereira, A. Plastino, B. Zadrozny, and L. Merschmann, "Categorizing feature selection methods for multi-label classification," *Artif. Intell. Rev.*, vol. 49, no. 1, pp. 57–78, Jan. 2018.

[6] S. Kashef, H. Nezamabadi-pour, and B. Nipour, "Multilabel feature selection: a comprehensiove review and guide experiments," *WIREs Data Min. Knowl. Discovery*, vol. 8, no. 2, p. e1240, Mar./Apr. 2018.

[7] J. Xu and Q. Ma, "Multi-label regularized quadratic programming feature selection algorithm with frank-wolfe method," *Expert Syst. Appl.*, vol. 95, pp. 14–31, Apr. 2018.

[8] W. Siblini, P. Kuntz, and F. Meyer, "A review on dimensionality reduction for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, pp. 1–20, in press, DOI:10.1109/TKDE.2019.2 940 014, 2019.

[9] R. Alhutaish, N. Omar, and S. Abdullah, "A comparison of multi-label feature selection methods using the algorithm adaptation approach," in *Proc. 4th Int. Visual Infor. Conf. (IVIC2015)*, ser. LNCS, vol. 9425, Bangi, Malaysia, Nov. 2015, pp. 199–212.

[10] D. Kong, C. H. Q. Ding, H. Huang, and H. Zhao, "Multi-label ReliefF and F-statistic feature selections for image annotation," in *Proc. 25th IEEE Conf. Comput. Vision Pattern Recognit.*, Providence, Rhode Island, USA, Jun. 2012, pp. 2352–2359.

[11] M.-L. Zhang, J. M. Pena, and V. Robles, "Feature selection for multi-label naive bayes classification," *Inf. Sci.*, vol. 179, no. 19, pp. 3218–3229, Sep. 2009.

[12] J. Yin, T. Tao, and J. Xu, "A multi-label feature selection algorithm based on multi-objective optimization," in *Proc. 27th Int. Joint Conf. Neural Networks (IJCNN2015)*, Killarney, Ireland, Jul. 2015, pp. 1–7.

[13] Q. Gao, Z. Li, and J. Han, "Correlated multi-label feature selection," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM2011)*, Glasgow, UK, Oct. 2011, pp. 1087–1096.

[14] Q. Xu, P. Zhu, Q. Hu, and C. Zhang, "Robust multi-label feature selection with missing labels," in *Proc. 7th Chin. Conf. Pattrern Recognit. (CCPR2016)*, ser. CCIS, vol. 662, Chengdu, China, Nov. 2016, pp. 752–765.

[15] J. Lee and D.-W. Kim, "Fast multi-label feature selection based on information-theoretic feature ranking," *Pattern Recognit.*, vol. 48, no. 9, pp. 2761–2771, Sep. 2015.

[16] S. Jungjit, A. A. Freitas, M. Michaelis, and J. Cinatl, "A multi-label correlation-based feature selection method for the classification of neuroblastoma microarray data," in *Proc. 12nd Ind. Conf. Data Min: Workshop Data Min. Life Sci. (ICDM2012)*, Berlin, Germany, Jul. 2012, pp. 149–157.

[17] J. Xu, "Effective and efficient multi-label feature selection approaches via modifying hilbert-schmidt independence criterion," in *Proc. 23rd Int. Conf. Neural Inf. Process. (ICONIP2016)*, ser. LNCS, vol. 9949, Kyoto, Japan, Nov. 2016, pp. 385–395.

[18] Y. Lin, Q. Hu, J. Liu, and J. Duan, "Multi-label feature selection based on max-dependency and min-redundancy," *Neurocompting*, vol. 168, pp. 92–103, Nov. 2015.

[19] J. Lee and D.-W. Kim, "SCLS: Multi-label feature selection based on scalable criterion for large label set," *Pattern Recognit.*, vol. 66, pp. 342–352, Jun. 2017.

[20] C. Liu, Q. Ma, and J. Xu, "Multi-label feature selection method combining unbiased hilbert-schmidt independence criterion with controlled genetic algorithm," in *Proc. 25th Int. Conf. Neural Inf. Process. (ICONIP2018)*, ser. LNCS, vol. 11304, Siem Reap, Cambodia, Dec. 2018, pp. 3–14.

[21] S. Jungjit and A. A. Freitas, "A new genetic algorithm for multi-label correlation-based feature selection," in *Proc. 23rd Eur. Symp. Artif. Neural Network, Artif. Intell. Mach. Learn. (ESANN2015)*, Bruges, Belgium, Apr. 2015, pp. 285–290.

[22] X. Wang, L. Zhao, and J. Xu, "Multi-label feature selection method based on multivariate mutual information and particle swarm optimization," in *Proc. 25th Int. Conf. Neural Inf. Process. (ICONIP2018)*, Seim Reap, Cambodia, Dec. 2018, pp. 84–95.

[23] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. 17th Int. Conf. on Mach. Learn. (ICML2000)*, Stanford, CA, USA, Jun.-Jul. 2000, pp. 359–366.

[24] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Proc. 4th Int. Workshop Nat. Inspired Cooperative Strategies Optim. (NICSO2010)*, Granada, Spain, May 2010, pp. 65–74.

[25] R. Y. M. Nakamura, L. A. M. Pereira, K. A. P. Costa, D. Rodrigues, J. P. Papa, and X. Yang, "BBA: a binary bat algorithm for feature selection," in *Proc. 25th SIBGRAPI Conf. Graphics, Pattern Image*, Ouro Preto, Brazil, Aug. 2012, pp. 291–297.

[26] X.-S. Yang, "Bat algorithm for multi-objective optimisation," *Int. J. Bio-inspired Comput.*, vol. 3, no. 5, pp. 267–274, 2011.

[27] L. Brezocnik, I. Fister, and V. Podgorelec, "Swarm intelligence algorithms for feature selection," *Appl. Sci*, vol. 8, no. 9, pp. Article–1521, Sep. 2018.

[28] I. Csiszar and J. Korner, *Information Theory*. Cambrige, UK: Cambridge University Press, 2011.

[29] A. Mariello and R. Battiti, "Feature selection based on the neighborhood entropy," *IEEE Tran. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6313–6322, Dec. 2018.

[30] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.