IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 26, 2019, accepted July 15, 2019, date of publication July 25, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2931035

Multi-Label Bioinformatics Data Classification With Ensemble Embedded Feature Selection

YUMENG GUO[1,2], FU-LAI CHUNG[2], GUOZHENG LI[1,3], AND LEI ZHANG[4]

[1]Department of Control Science and Engineering, Tongji University, Shanghai 201804, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[3]Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China
[4]Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

Corresponding author: Guozheng Li (drgzli@gmail.com)

This work was supported in part by the GRF, UGC under project Hong Kong PolyU, under Grant 152039/14E, in part by the Central Research Grant, Hong Kong PolyU, under Project G-YBVT, in part by the International Exchange Program for Graduate Students, Tongji University, under Grant 2017010007, in part by the National Key Research and Development Program of China under Grant 2017YFC1703501, in part by the Fundamental Research Funds for the Central Public Welfare Research Institutes under Grant ZZ0908032, and in part by the National Natural Science Foundation of China under Grant 81503680.

ABSTRACT In bioinformatics, the vast of multi-label type of datasets, including clinical text, gene, and protein data, need to be categorized. Specifically, due to the redundant or irrelevant features in bioinformatics data, the performance of multi-label classifiers will be limited, and therefore, selecting effective features from the feature space is necessary. However, most of the proposed methods, which aimed at dealing with multi-label feature selection problem in the past few years, only adopt a simple and direct strategy that transforms the multi-label feature selection problem into more single-label ones and ignore correlations among different labels. In this paper, a novel algorithm named ensemble embedded feature selection (EEFS) is proposed to handle multi-label bioinformatics data learning problem in a more effective and efficient way. The EEFS does not only explicitly find out the correlations among labels, but it can also adequately utilize the label correlations by multi-label classifiers and evaluation measures. Furthermore, it can reduce the accumulated errors of data itself by employing an ensemble method. The experimental results on five multi-label bioinformatics datasets show that our algorithm achieves significant superiority over the other state-of-the-art algorithms.

INDEX TERMS Bioinformatics, multi-label learning, embedded feature selection.

I. INTRODUCTION

Multi-label type of bioinformatics data widely exists in clinical text data [1], gene data [2], protein data [3], [4] and so on. For example, a patient suffering from *cough* and *fever* should be associated with both two disease labels in the clinical records. Formally, let $\mathcal{X} = \mathbb{R}^d$ denote the $d-$dimensional feature space and $\mathcal{Y} = \{0, 1\}^q$ denote the $q-$dimensional label space, where each example in multi-label bioinformatics data can be denoted as $(\boldsymbol{x}_i, Y_i)(\boldsymbol{x}_i \in \mathcal{X}, Y_i \subseteq \mathcal{Y})$. Due to the rapidly expanding quantity of multi-label bioinformatics data resources, techniques based on multi-label learning demonstrate the superiority in mining useful information from vast this type of data [5], [6]. These techniques aim to build classification models for instances which are assigned with multiple labels simultaneously. However, the feature space of multi-label data inevitably exists redundant and irrelevant features which could limit the performance of multi-label classifiers. Encouragingly, to promote the classification performance, many multi-label feature learning methods, which could reduce the dimension of feature space, have been proposed to acquire important and effective information from the original feature space [7], [8]. Specifically, there are two kinds of dimension reduction methods for multi-label data: feature extraction and feature selection. For feature extraction, unsupervised approaches, such as principal component analysis (PCA) [9], latent semantic indexing (LSI) [10], [11], multi-output regularized feature projection [12], extraction shared subspaces [13], are proposed to find a compact feature space to represent the original datasets and supervised approaches, such as linear discriminant analysis (LDA) [14]–[16] and multi-label dimensionality reduction via dependence maximization which is based on the Hilbert-Schmidt independence criterion (MDDM) [17], achieve better performance. These approaches are effective to improve the performance of classification. However, the

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Rathinam.

extracted features fuse the information of original features, and lose the distinct physical meanings. Hence, the features extracted from the original feature space can be hardly explained and easily comprehended.

These characteristics limit the use of feature extraction methods in particular research area, such as multi-label bioinformatics data analysis. For example, many clinical decision-making tasks following the philosophy of Evidence Based Medicine (EBM) rely on the ability to find relevant health records and gather sufficient clinical evidence [18]. Therefore, reliable feature subsets which have interpretation of physical significance for clinical records classification can help doctors or researchers in disease diagnosis, prevention and treatment, or simply medical text resources categorization and retrieving. Encouragingly, different from feature extraction, feature selection approaches remains the physical meaning of features when reducing the feature dimension. Hence, it is more suitable than feature extraction for dealing with this type of multi-label data analysis task.

To deal with the multi-label feature selection task, there are three main types of methods: filter, wrapper ands embedded [19]–[21]. We will describe them in detail in the next section. In this paper, a new embedded methods type of multi-label feature selection algorithm named EEFS, i.e. *Ensemble Embedded Feature Selection*, for multi-label bioinformatics data is proposed. It randomly selects partial training examples to train classification models which is an ensemble method, and then employs evaluation measure and averaged training examples for each column to test the trained models iteratively to acquire the final feature importance ranking. Experiment results demonstrate our algorithm is superior over other multi-label feature selection (feature importance ranking) algorithms. This paper extends our preliminary work [22].

The rest of this paper is organized as follows. Section II, reviews the existing multi-label feature selection methods. Section III, presents the proposed EEFS algorithm. Section IV, presents the design of the experiments. Section V, reports and analyzes the comparative experimental results. Finally, Section VI, summarizes several issues and suggest some future directions.

II. RELATED WORKS

Recently, multi-label feature selection methods for bioinformatics data have received increasing attention from research community, due to the rapidly expanding quantity of multi-label bioinformatics data resources. There is a rich body of work on the research of them. As mentioned above, the existing multi-label feature selection methods can be generally categorized into three classes, namely filter, wrapper and embedded methods. In this section, we will review the main algorithms of these three main types of methods in detail.

- Filter methods
  The main idea of filter feature selection methods [23], [24] for multi-label bioinformatics classification is transforming the single-label methods to multi-label methods. For example, Yang and Pedersen [23] proposed a filter framework to evaluate features for each label separately under some statistic evaluation measures, and combine the results by the maximal or average methods. This framework is an extension of single-label filter feature selection methods. It deals with the labels separately, which ignores the correlations within labels. Let $\mathcal{X} = \{f_1, f_2, \ldots, f_d\}$ denote the feature space with d features and $\mathcal{Y} = \{l_1, l_2, \ldots, l_q\}$ denote the label space with q class labels. Then the maximal and average types of filter multi-label feature selection methods, FS_{max} and FS_{avg}, are defined as follows:

$$FS_{max}(f_i) = \max_q \{EM(f_i, l_1), \ldots, EM(f_i, l_q)\} \quad (1)$$

$$FS_{avg}(f_i) = \frac{1}{q}\sum_{j=1}^{q} EM(f_i, l_j) \quad (2)$$

  where EM is the evaluation measure which evaluates the correlations between feature and label for single-label feature selection. The evaluation measures utilized by single-label feature selection methods can be χ^2, Relief [25], COR [26] and mRMR [27]. The importance which is represented by the value $FS(f_i)$ of ith features decided in multi-label bioinformatics data depends on the rules of filter multi-label feature selection as shown in Eq. (1) or Eq. (2). The results of filter multi-label feature selection will demonstrate the feature importance ranking. These methods have linear computation cost, but their selection results are always rough. They consider the relevance between labels and each feature, while ignoring the power when features combined together. Moreover, filter methods provide a unique feature ranking for different kind of classifier. The selected feature subset is always not the most suitable subset for a certain classifier.
- Wrapper methods
  The main idea of wrapper feature selection method [28]–[30] for multi-label learning is depending on the learning machine and utilizing the learning machine of interest as a black box to score feature subsets according to their predictive power. They are widely used in scientific data analysis, because the selected feature subset is optimal to the specific learning machine due to its mechanism that the selection result is based on the learning algorithms. For example, Shao *et al.* [30] propose a hybrid optimization multi-label feature selection method called HOML. In their work, simulated annealing, genetic algorithm and hill climb strategies are combined to generate many feature subsets and then they utilize multi-label classifiers to select the best one. The process of generating and selecting optimal feature subset is comparative time-consuming. These methods are classifier specified feature selection methods. Specifically, they select a wide variety of feature subsets based on some principle from training data to train corresponding classifiers, and then

measure the selected feature subsets from test data with the corresponding trained classifiers directly. Wrapper methods can improve the performance of classifiers in a large range. However, their computational complexity is always too high.

- Embedded methods
  The embedded feature selection methods [31]–[33] for multi-label learning are a trade-off way to overcome the weaknesses of filter and wrapper methods. For embedded methods, the multi-label classifiers are embedded in the process of feature selection to compute the relationships between features and classifiers. It also can avoid the time consuming problem compared with wrapper methods. For example, You *et al.* [31] proposed an algorithm named multi-label embedded feature selection (MEFS) which utilizes the prediction risk and classifier to evaluate the features importance in feature subset and backward search strategy to select the best feature from feature subset step by step. With this, the selected features are more directly to improve the classification performance. But with the change of the training data, the trained model will generate different feature rankings, so it is difficult to get relatively stable feature ranking. This will not be conductive to further analysis of the data and the generalization of the algorithm. Furthermore, MEFS cannot overcome the weakness of reduce the accumulated errors of data itself.

In this paper, due to the proposed EEFS algorithm providing the feature importance ranking as final result which is similar as algorithms based on filter methods and different from algorithms based on wrapper methods only providing optimal feature subset from the candidate feature subsets, we only compare our algorithm with algorithms based on filter and embedded methods. And we do not involve algorithms based on wrapper methods in the experimental part.

III. THE PROPOSED ALGORITHM

In this section, our proposed algorithm EEFS, i.e. *Ensemble Embedded Feature Selection*, will be presented for multi-label bioinformatics data feature selection. EEFS can provide relatively stable feature ranking and reduce the negative effect of the change of training data. Furthermore, it can be adjusted on the basis of the multi-label bioinformatics data structural characteristics for boosting the performance of multi-label classifiers. Specifically, aimed at generating the feature subset which can be utilized to improve the classifier's performance, EEFS employs prediction risk and forward search strategy to evaluate the importance of features. And EEFS's feature selection process cooperates with multi-label classifier and prediction risk. In detail, the feature selection capacity of EEFS mainly relies on the classifiers' learning ability and the employed evaluation measure which is used for computing prediction risk.

Prediction risk can evaluate the models' classification performance. During the learning process of models, prediction risk is applied to estimate the prediction accuracy of the models and then select suitable models. The principle of prediction risk minimization is often used for selecting optimal feature subset in single-label problems. Prediction risk criteria estimates each feature by computing the difference between original and updated training data's results which are from testing the trained model. The updated training data means the value of a certain feature for each training example is replaced by its mean value of all training examples. The prediction risk (*preRISK*) of *i*th feature is defined as follows:

$$preRISK = preERR(\overline{\boldsymbol{x}}^i) - preERR \tag{3}$$

where *preERR* stands for the prediction error of trained model on original training data and $preERR(\overline{\boldsymbol{x}}^i)$ stands for the prediction error of trained model on updated training data corresponding to *i*th feature. In experimental part, we employ *average precision*, which is a multi-label ranking type of evaluation measure, to compute *preERR* and it is defined as follows:

$$average\ precision = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i|}\sum_{l_k\in Y_i}\frac{|\mathcal{R}(\boldsymbol{x}_i, l_k)|}{rank(\boldsymbol{x}_i, l_k)}$$

Algorithm 1 The EEFS Algorithm

Inputs:

$\mathcal{D}$: bioinformatics training data $\{(\boldsymbol{x}_i, Y_i)|1 \le i \le n\}$ $(\boldsymbol{x}_i \in \mathcal{X}, Y_i \subseteq \mathcal{Y}, \mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{l_1, l_2, \ldots, l_q\})$

$\mathfrak{L}$: the loss function in Eq. (4)

μ: the percentage parameter $(0 < \mu \le 100\%)$

λ: the iteration times (any integer ≥ 1)

Outputs:

r: the feature ranking list

$r \leftarrow \emptyset$//empty feature ranking list
$\boldsymbol{u} \leftarrow [1, 2, \ldots, d]$//$\boldsymbol{u}$ is the complete feature set, initialize it by the original set
$preRISK \leftarrow (0, \ldots, 0)$ with the dimensionality $|\boldsymbol{u}|$ //initialize *preRISK*
for $j = 1$ to λ **do**
  $\mathcal{D}_r = (\boldsymbol{x}^r, Y^r)$//randomly select $(\mu \cdot n)$ examples form training data $\mathcal{D}$
  $model \leftarrow$ train classifier$(\boldsymbol{x}^r, Y^r)$//train a classifier with the randomly selected data
  $preERR \leftarrow$ test classifier$(model, \boldsymbol{x}^r, Y^r)$//test the trained classifier and get the $\mathcal{D}^r$ error
  for $k = 1$ to d **do**
    $preERR(\overline{\boldsymbol{x}^r}^k) \leftarrow$ test classifier$(model, \overline{\boldsymbol{x}^r}^k, Y^r)$ //test the trained classifier and get $\overline{\boldsymbol{x}^r}^k$ error
    compute $preRISK_j^k$ according to Eq. (3)
  end for// evaluate each feature's importance according to the prediction risk criterion
end for
compute $\overline{preRISK}_j^k$ for each feature according to Eq. (5)
$r \leftarrow rank[\overline{preRISK}_j^k]$//update the feature ranking list
output the final feature ranking list r

TABLE 1. Comparison of the optimal results of Four algorithms with BR classifier on Five datasets.

Dataset	Algorithm	*mlACC*	*mlPRE*	*mlREC*	*mlF1*	*ACC*
clinical	EEFS	**0.8325±0.0195**	**0.8691±0.0196**	**0.8868±0.0187**	**0.8758±0.0150**	**0.7586±0.0282**
	MEFS	0.7990±0.0195	0.8578±0.0196	0.8724±0.0187	0.8513±0.0150	0.7146±0.0282
	max	0.8106±0.0231	0.8538±0.0162	0.8799±0.0205	0.8589±0.0136	0.7350±0.0312
	avg	0.8078±0.0252	0.8529±0.0189	0.8799±0.0242	0.8551±0.0192	0.7254±0.0312
	benchmark	0.7990±0.0195	0.8273±0.0196	0.8724±0.0187	0.8490±0.0150	0.6980±0.0282
medical	EEFS	0.7511±0.0404	0.8190±0.0240	**0.8405±0.0519**	0.8278±0.0433	0.6942±0.0507
	MEFS	0.7511±0.0402	0.8190±0.0240	0.8395±0.0497	0.8278±0.0377	0.6942±0.0446
	max	0.7525±0.0468	**0.8268±0.0360**	0.8376±0.0568	0.8309±0.0428	0.6994±0.0513
	avg	**0.7533±0.0476**	0.8263±0.0401	0.8376±0.0526	**0.8315±0.0439**	**0.7004±0.0568**
	benchmark	0.7511±0.0573	0.8190±0.0240	0.8376±0.0654	0.8278±0.0397	0.6942±0.0505
plant	EEFS	**0.7553±0.0249**	**0.8756±0.0241**	**0.7590±0.0299**	**0.8098±0.0207**	**0.7203±0.0230**
	MEFS	0.7506±0.0304	0.8672±0.0275	0.7515±0.0413	0.8023±0.0251	0.7162±0.0270
	max	0.7523±0.0297	0.8679±0.0234	0.7556±0.0397	0.8063±0.0216	0.7172±0.0225
	avg	0.7539±0.0207	0.8710±0.0261	0.7564±0.0257	0.8070±0.0195	0.7193±0.0197
	benchmark	0.7465±0.0226	0.8672±0.0275	0.7467±0.0282	0.8020±0.0198	0.7141±0.0181
virus	EEFS	**0.8583±0.0591**	0.9012±0.0556	**0.8857±0.0565**	**0.8928±0.0503**	0.8114±0.0683
	MEFS	0.8577±0.0602	**0.9085±0.0394**	0.8750±0.0597	0.8847±0.0482	0.8114±0.0683
	max	0.8577±0.0602	**0.9085±0.0523**	0.8750±0.0597	0.8875±0.0410	0.8114±0.0683
	avg	0.8577±0.0602	**0.9085±0.0523**	0.8750±0.0597	0.8875±0.0410	0.8114±0.0683
	benchmark	0.8577±0.0602	0.8965±0.0529	0.8750±0.0597	0.8847±0.0482	0.8114±0.0683
yeast	EEFS	**0.5079±0.0165**	0.7091±0.0205	**0.5896±0.0119**	**0.6417±0.0115**	**0.1593±0.0179**
	MEFS	0.5011±0.0140	0.7094±0.0200	0.5828±0.0083	0.6398±0.0114	0.1527±0.0191
	max	0.5011±0.0140	**0.7116±0.0195**	0.5762±0.0079	0.6365±0.0110	0.1510±0.0181
	avg	0.5024±0.0140	0.7114±0.0195	0.5769±0.0080	0.6357±0.0109	0.1523±0.0184
	benchmark	0.5011±0.0140	0.7091±0.0205	0.5762±0.0079	0.6357±0.0109	0.1407±0.0165

* benchmark: It is not an algorithm and represents the cross validation classification results of corresponding multi-label classifiers and evaluation measures about the related datasets with all the features (no feature selection).

where

$$\mathcal{R}(\boldsymbol{x}_i, l_k) = \{l_j | rank(\boldsymbol{x}_i, l_j) \leq rank(\boldsymbol{x}_i, l_k), l_j \in Y_i\},$$

$rank(\boldsymbol{x}, l)$ returns the rank of $l \in Y$ based on the descending order induced from model and n is the number of examples. *Average precision* evaluates the average fraction of relevant labels ranked higher than a particular label $l_k \in Y_i$.

The *preRISK* of ith feature captures the *preERR* difference between the updated training data which replaces the ith feature's value for each example by its mean value of all examples and the original training data, when they test the trained model separately. We employ the *preRISK* value as the ranking basis of feature importance. When we utilize the prediction risk to reduce the dimension of feature space in multi-label bioinformatics data, we employ the evaluation measures of multi-label learning as the loss function for prediction risk. $x_j^i \in \mathbb{R}$ is the value of ith feature for jth example. The output of a classifier $\mathcal{C}(\boldsymbol{x})$ ($\boldsymbol{x} = [x^1, \ldots, x^d]$) is the predicted label sets Y'. Let $\mathfrak{L}(Y', Y)$ denote a multi-label loss function where Y is the true label set associated with instance $\boldsymbol{x}$. Then $preERR(\overline{\boldsymbol{x}}^i)$ is defined as follows:

$$preERR(\overline{\boldsymbol{x}}^i) = \mathfrak{L}(\mathcal{C}([x^1, \ldots, \overline{x}^i, \ldots, x^d]), Y) \tag{4}$$

where $\overline{x}^i$ is the mean value of the ith feature of all examples and $\mathcal{C}([x^1, \ldots, \overline{x}^i, \ldots, x^d])$ is the prediction value of all examples with the ith feature replaced by their mean value.

To further improve the performance of EEFS, we randomly select μ ($0 < \mu \leq 100\%$) percentage of instances from the original training data to train models, and then utilize Eq. (4) to compute the prediction risk for each feature. At last, we repeat this process for λ (In theory, any integer ≥ 1; In practice, $1 \leq \lambda \leq 20$ on the basis of our experimental experience) times. In the jth ($1 \leq j \leq \lambda$) iteration, we compute the prediction risk of the ith feature $preRISK_j^i$ according to Eq. (3). The average prediction risk $\overline{preRISK}^i$ of ith feature for all λ iterations is computed as follows:

$$\overline{preRISK}^i = \frac{1}{\lambda} \sum_{j=1}^{\lambda} preRISK_j^i \tag{5}$$

In order to better select appropriate parameters μ and λ in the multi-label bioinformatics feature selection process, we found a knack of good guiding significance according to our experimental experience. Firstly, we employ other state-of-the-art multi-label feature selection algorithms (e.g. based on filter methods) as benchmark to compute the importance of each feature. Secondly, we use EEFS with setting:

TABLE 2. Comparison of the optimal results of four algorithms with CC classifier on five datasets.

Dataset	Algorithm	*mlACC*	*mlPRE*	*mlREC*	*mlF1*	*ACC*
clinical	EEFS	**0.8521±0.0255**	**0.8511±0.0343**	**0.8896±0.0160**	**0.8697±0.0224**	**0.7759±0.0427**
	MEFS	0.8391±0.0255	0.8379±0.0343	0.8794±0.0160	0.8580±0.0224	0.7599±0.0427
	max	0.8320±0.0156	0.8365±0.0221	0.8712±0.0154	0.8512±0.0142	0.7573±0.0250
	avg	0.8349±0.0165	0.8382±0.0235	0.8750±0.0116	0.8536±0.0152	0.7592±0.0290
	benchmark	0.8174±0.0255	0.8193±0.0343	0.8574±0.0160	0.8378±0.0224	0.7286±0.0427
medical	EEFS	**0.7503±0.0481**	**0.8194±0.0283**	**0.8042±0.0274**	**0.8104±0.0390**	0.7126±0.0740
	MEFS	0.7455±0.0426	0.8169±0.0283	0.7991±0.0341	0.8075±0.0373	0.7045±0.0543
	max	0.7473±0.0414	0.8190±0.0283	0.7996±0.0335	0.8087±0.0272	**0.7127±0.0529**
	avg	0.7455±0.0374	0.8111±0.0283	0.7988±0.0387	0.8024±0.0279	0.7065±0.0480
	benchmark	0.7455±0.0429	0.8046±0.0283	0.7979±0.0526	0.8009±0.0334	0.7065±0.0502
plant	EEFS	0.7833±0.0197	0.8046±0.0198	0.7804±0.0298	0.7921±0.0216	0.7502±0.0210
	MEFS	0.7852±0.0251	0.8089±0.0309	0.7785±0.0296	0.7926±0.0265	0.7565±0.0233
	max	0.7912±0.0270	0.8109±0.0285	0.7879±0.0338	0.7990±0.0281	0.7575±0.0260
	avg	**0.7970±0.0204**	**0.8206±0.0267**	**0.7909±0.0290**	**0.8052±0.0235**	**0.7637±0.0232**
	benchmark	0.7766±0.0230	0.8001±0.0317	0.7726±0.0279	0.7859±0.0267	0.7430±0.0190
virus	EEFS	0.8535±0.0490	**0.9044±0.0594**	0.8574±0.0667	0.8781±0.0419	0.8117±0.0544
	MEFS	0.8534±0.0508	0.8967±0.0629	0.8553±0.0662	0.8735±0.0462	0.8017±0.0650
	max	**0.8720±0.0507**	0.8967±0.0629	**0.8788±0.0589**	**0.8857±0.0478**	**0.8212±0.0703**
	avg	**0.8720±0.0507**	0.8967±0.0629	**0.8788±0.0589**	**0.8857±0.0478**	**0.8212±0.0703**
	benchmark	0.8534±0.0508	0.8967±0.0629	0.8553±0.0662	0.8735±0.0462	0.8017±0.0650
yeast	EEFS	**0.4597±0.0083**	**0.5970±0.0158**	**0.5636±0.0195**	**0.5798±0.0100**	**0.1721±0.0190**
	MEFS	0.4581±0.0112	0.5964±0.0327	0.5635±0.0170	0.5795±0.0100	0.1700±0.0194
	max	0.4457±0.0094	0.5835±0.0191	0.5509±0.0200	0.5667±0.0119	0.1634±0.0200
	avg	0.4380±0.0101	0.5766±0.0172	0.5445±0.0186	0.5600±0.0097	0.1552±0.0179
	benchmark	0.4225±0.0089	0.5627±0.0131	0.5319±0.0180	0.5468±0.0112	0.1427±0.0191

* benchmark: It is not an algorithm and represents the cross validation classification results of corresponding multi-label classifiers and evaluation measures about the related datasets with all the features (no feature selection).

$\mu = 100\%$ and $\lambda = 1$ to estimate the importance of each feature. After that the features are ranked according to their importance in descending order. Thirdly, we select the same top percentage of features from the two feature rankings as feature subsets, and denote them as subsets A and B respectively, and then utilize multi-label classifiers to test their performance. Then we set the parameters by comparing the performance of the two feature subsets. If the performance of A is better than B, we set μ close to 100% and λ small. On the contrary, if the performance of B is better than A, we set μ away from 100% and λ big. Because this method essentially reduces the accumulated errors of data itself by employing ensemble method, it can acquire satisfactory experimental results. The pseudo code of EEFS is demonstrated in Algorithm 1.

IV. EXPERIMENTS

A. EXPERIMENTAL DATASETS

Five bioinformatics datasets are employed to help compare our proposed algorithm with the other state-of-the-art algorithms. For each dataset $\mathcal{S} = \{(\boldsymbol{x}_i, Y_i)|1 \leq i \leq p\}$, we use $|\mathcal{S}|$, $dim(\mathcal{S})$, $L(\mathcal{S})$ and $F(\mathcal{S})$ to denote the *number of examples, number of features, number of class labels, and feature type* for $|\mathcal{S}|$ respectively. The *clinical* dataset [34], [35] comprises a total number of 1566 free text clinical records label by disease codes. The content of the records is mostly composed of patient's impressions reported by some radiologists in free text form. We extract bag-of-words features from the raw text and further transform word counts into TF-IDF features. Only the word frequencies of the top 232 words are kept after stop words filtering and word stemming [36]. The disease labels are expressed by a group of ICD-9-CM codes [37]. It contains a list of carefully categorized disease entries, coded by distinguished numbers, which can be used to classify the clinical records into their relevant diseases. In our experiments, we restrict the label size to top 10 for analysis and algorithm comparisons. Also as clinical text data, the data processing mode of *medical* dataset [34] is similar to *clinical*. It comprises 978 text medical records, 217 top frequencies words, and 20 labels. Two protein datasets, which are *plant* [38], [39] and *virus* [40], [41], with experimentally determined subcellular location are obtained from Cell-Ploc 2.0 [42]. In these two datasets, protein sequences were totally collected from the Swiss-Prot database at http://www.ebi.ac.uk/swissprot/. We use *go* protein representation method, which is widely used in many existing protein subcellular localization systems, to generate features of protein examples [43]–[45]. Information of these two

TABLE 3. Comparison of the optimal results of four algorithms with MLkNN classifier on five datasets.

Dataset	Algorithm	*mlACC*	*mlPRE*	*mlREC*	*mlF1*	*ACC*
clinical	EEFS	0.7389±0.0321	**0.8632±0.0384**	**0.7383±0.0301**	**0.7952±0.0254**	**0.6647±0.0399**
	MEFS	0.6995±0.0212	0.8397±0.0315	0.6901±0.0271	0.7566± 0.0147	0.6226±0.0281
	max	**0.7418±0.0308**	0.8460±0.0237	0.7367±0.0358	0.7871±0.0251	**0.6647±0.0406**
	avg	0.7077±0.0333	0.8373±0.0298	0.7123±0.0370	0.7691±0.0247	0.6315±0.0332
	benchmark	0.5889±0.0279	0.7638±0.0350	0.5842±0.0268	0.6618±0.0269	0.5057±0.0329
medical	EEFS	**0.7157±0.0468**	**0.8422±0.0305**	**0.7379±0.0491**	**0.7837±0.0375**	**0.6738±0.0561**
	MEFS	0.6796±0.0499	0.8347±0.0289	0.7166±0.0405	0.7705±0.0364	0.6472±0.0460
	max	0.6925±0.0501	0.8365±0.0582	0.7183±0.0576	0.7727±0.0515	0.6553±0.0540
	avg	0.6822±0.0480	0.8333±0.0445	0.7062±0.0598	0.7643±0.0475	0.6421±0.0480
	benchmark	0.6119±0.0571	0.7893±0.0427	0.6432±0.0623	0.7083±0.0487	0.5592±0.0645
plant	EEFS	**0.7492±0.0332**	**0.8530±0.0450**	0.7442±0.0302	**0.7940±0.0245**	**0.7079±0.0353**
	MEFS	0.7448±0.0216	0.8398±0.0342	**0.7455±0.0302**	0.7892±0.0207	0.7039±0.0217
	max	0.7061±0.0462	0.8330±0.0230	0.6974±0.0518	0.7581±0.0304	0.6739±0.0362
	avg	0.6988±0.0270	0.8497±0.0391	0.6923±0.0325	0.7620±0.0216	0.6656±0.0317
	benchmark	0.6302±0.0426	0.8153±0.0363	0.6244±0.0395	0.7063±0.0295	0.5924±0.0465
virus	EEFS	**0.7853±0.1059**	**0.9058±0.0763**	**0.7978±0.1109**	**0.8367±0.0825**	**0.7336±0.1265**
	MEFS	0.7777±0.0555	0.8875±0.0836	0.7973±0.0643	0.8362±0.0442	0.7145±0.0827
	max	0.7752±0.0695	0.8762±0.0617	0.7866±0.0691	0.8285±0.0623	0.7098±0.0879
	avg	0.7752±0.0695	0.8762±0.0617	0.7866±0.0691	0.8285±0.0623	0.7098±0.0879
	benchmark	0.7477±0.0764	0.8582±0.0477	0.7554±0.0732	0.8023±0.0554	0.6710±0.0831
yeast	EEFS	**0.5186±0.0188**	0.7285±0.0172	0.5889±0.0192	**0.6488±0.0180**	**0.1920±0.0254**
	MEFS	0.5170±0.0112	0.7286±0.0255	0.5889±0.0192	0.6485±0.0179	0.1837±0.0220
	max	0.5170±0.0183	0.7289±0.0219	0.5889±0.0192	0.6485±0.0179	0.1845±0.0224
	avg	0.5170±0.0183	**0.7303±0.0199**	0.5889±0.0192	0.6485±0.0179	0.1907±0.0271
	benchmark	0.5170±0.0183	0.7219±0.0227	0.5889±0.0192	0.6485±0.0179	0.1820±0.0220

* benchmark: It is not an algorithm and represents the cross validation classification results of corresponding multi-label classifiers and evaluation measures about the related datasets with all the features (no feature selection).

protein datasets are described as follows: 1) 969 different proteins with 224 *go* features distributed among 12 subcellular location for *plant* cells; 2) 206 different proteins with 185 *go* features distributed among 6 subcellular location for *virus* cells. The *yeast* dataset [2] is formed by micro-array expression data and phylogenetic profiles with 2417 genes. The input dimension is 103. Each gene is associated with a set of functional labels whose size can be 14. Table 4 briefly demonstrates the characteristics of the experimental datasets.

B. MULTI-LABEL FEATURE SELECTION ALGORITHMS

We compare our proposed algorithm with the following three algorithms: Multi-label embedded feature selection (MEFS) [31], χ^2 based maximal and average type of filter multi-label feature selection (*max* and *avg*) [23].

- MEFS: The basic idea of this algorithm is to utilize the prediction risk and classifier to evaluate the features importance in feature subset and backward search strategy to select the best feature form feature subset step by step to form feature ranking.
- *max*: The basic idea of this algorithm is to calculate the dependency score with a χ^2 based evaluation statistic between a feature and a label separately. The maximal dependency score of a certain feature across all labels stands for the final importance score of this feature. According to the importance score of each feature, we get feature ranking in descending order.
- *avg*: The basic idea of *avg* is similar to *max*. Dependency scores for a certain feature on all labels are averaged to form the final importance score for this feature.

C. MULTI-LABEL CLASSIFIERS

In order to eliminate the bias of classifiers, three multi-label classifiers, which are *Binary Relevance* (BR) [46], *Classifier Chain* (CC) [47] and *Multi-Label k-Nearest Neighbor* (MLkNN) [48] are employed in the experiment. BR and CC are problem transformation method, which transform the multi-label classification problem into one or more single-label classification. MLkNN is algorithm adaptation method, which extends specific learning algorithms in single label problem to handle multi-label data directly.

- *Binary Relevance* (BR): The basic idea of this algorithm is to decompose the multi-label learning problem into q independent binary classification problems, where each binary classification problem corresponds to a possible label in the label space. In brief, it trains and tests models for each label.

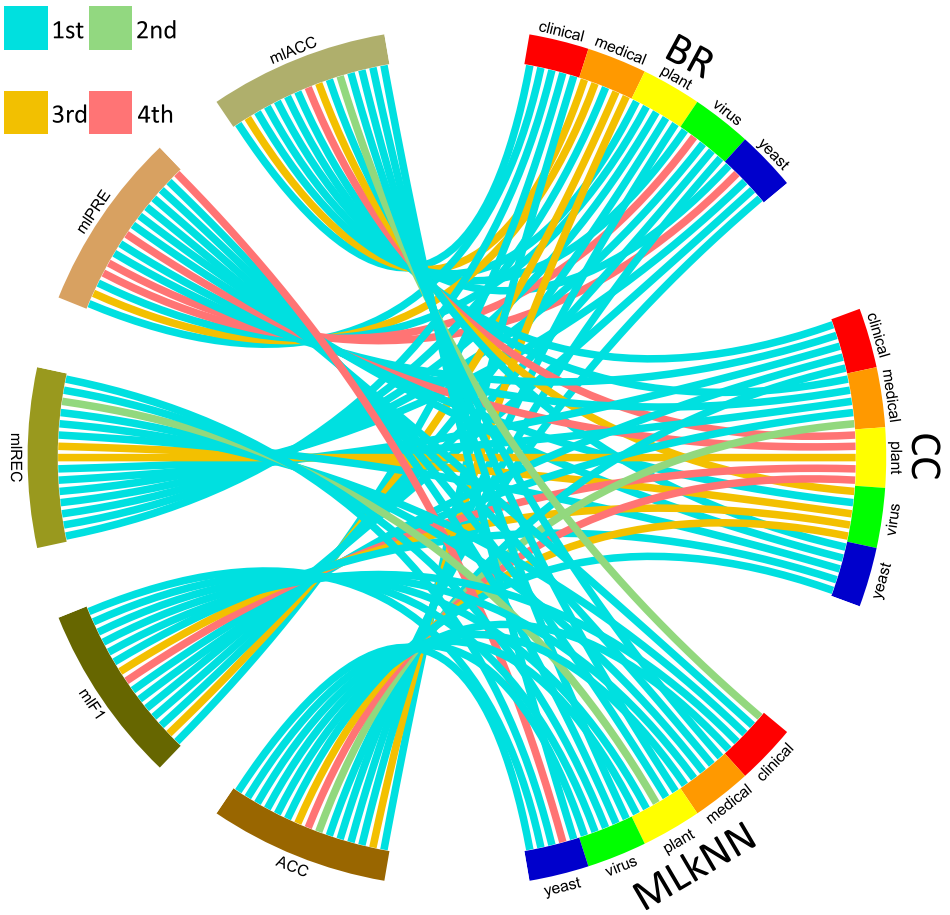


FIGURE 1. Best performance of EEFS with BR, CC and MLkNN classifiers on five datasets compared with other three algorithms. (Each dataset connects all the evaluation measures with different color curves simultaneously and the number of color curves on the left half circle denotes the rank performance of EEFS corresponding to each evaluation measure).

- *Classifier Chain* (CC): The basic idea of classifier chain is to transform the multi-label learning problem into a chain of binary classification problems, where subsequent binary classifiers in the chain is built upon the predictions of preceding ones. In brief, it treats label as new feature with original feature space to predict next label in a chain way.
- *Multi-Label k-Nearest Neighbor* (MLkNN). The basic idea of this algorithm is adapting k-nearest neighbor techniques to deal with multi-label data, where maximum a posteriori (MAP) rule is utilized to make prediction by reasoning with the labeling information embodied in the neighbors. In brief, it designs a new algorithm based on k-nearest neighbor techniques for multi-label data.

D. EVALUATION MEASURES

In the multi-label learning community, it is well known that the performance evaluation of multi-label learning differs from that of classical single-label learning because each example could have multiple labels simultaneously. Therefore five standard evaluation measures, which are *multi-label accuracy (mlACC), precision (mlPRE), recall (mlREC), F1 (mlF1)* and *subset accuracy (ACC)*, are introduced for evaluating the performance of our proposed method from multiple aspects more exactly [49], [50]. The five evaluation measures are defined as follows:

$$mlACC = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \bigcap Y_i'|}{|Y_i \bigcup Y_i'|}$$

$$mlPRE = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \bigcap Y_i'|}{|Y_i'|}$$

$$mlREC = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \bigcap Y_i'|}{|Y_i|}$$

$$mlF1 = \frac{2 \cdot mlPRE \cdot mlREC}{mlPRE + mlREC}$$

$$ACC = \frac{1}{m}\sum_{i=1}^{p} 1(Y_i \equiv Y_i')$$

where m is the number of test examples, Y_i and Y_i' are the set of true labels and the set of predicted labels of each instance, respectively. $mlF1$ is the harmonic mean of $mlREC$

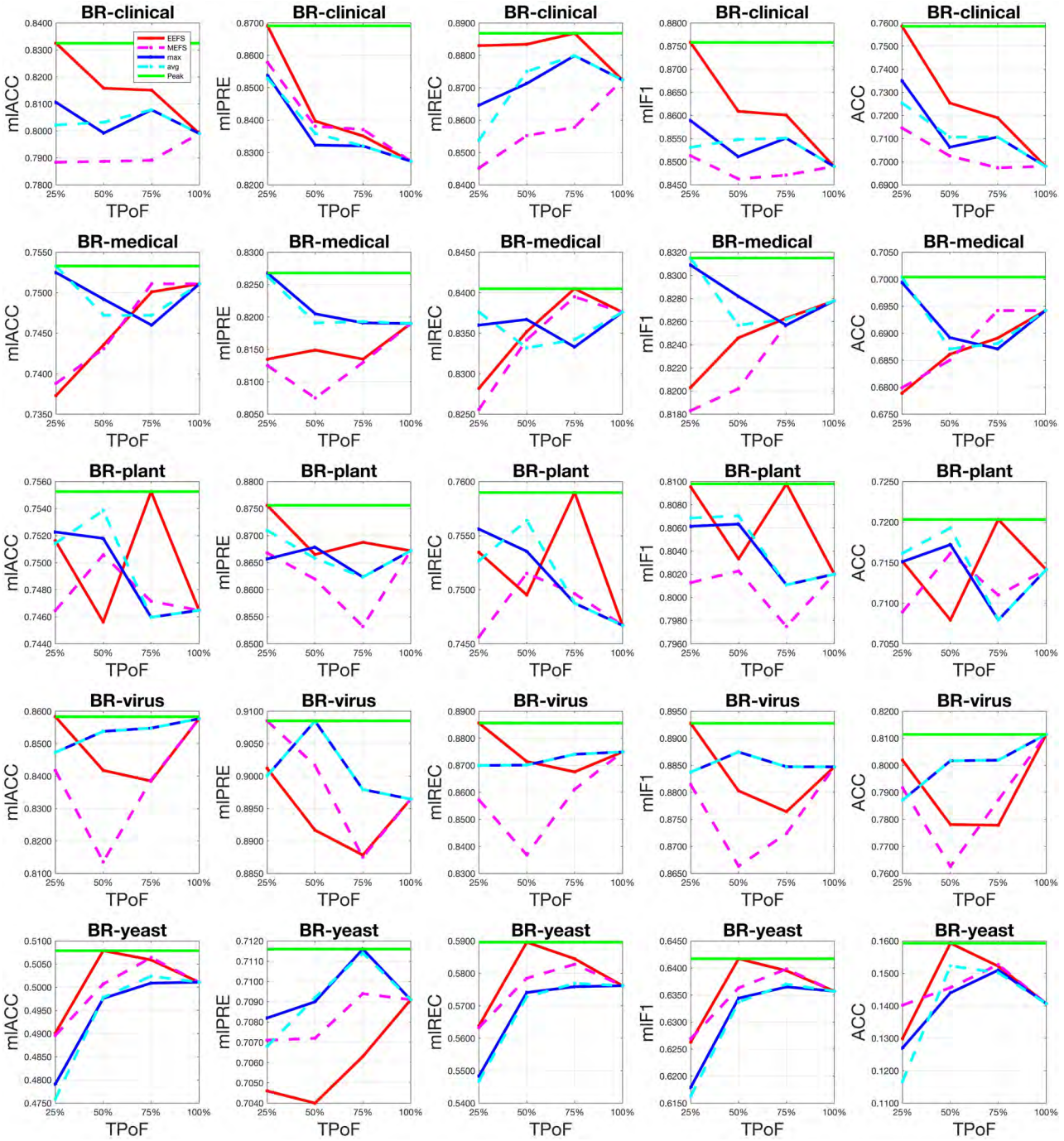


FIGURE 2. Performance of the four algorithms with BR classifier on five datasets (TPoF: Top percentage of feature ranking in descending order according to their importance).

and *mlPRE*. For the five evaluation measures, note that the bigger the measure value, the better the performance.

E. EXPERIMENT CONFIGURATION

In the experiments, multi-label ranking type of evaluation measure *average precision* [51], [52] is employed as *preERR* to compute prediction risk. EEFS is compared with 3 other state-of-the-art feature selection methods MEFS, *max* and *avg*. 3 classifiers and 5 evaluation measures previously described are all implemented in the experiment for an exhaustive assessment. In the experiment, we select top 25%, 50%, 75% and 100% percentage of the features, which are ranked in a descending order according to their importance, to demonstrate the results and we employ 10-fold cross validation in the experimental part. For the setting of parameters μ and λ of EEFS, we set $\mu = 80\%$ and $\lambda = 5$ for BR and CC and $\mu = 50\%$ and $\lambda = 15$ for MLkNN according to the previous analysis of experimental experience we mentioned.

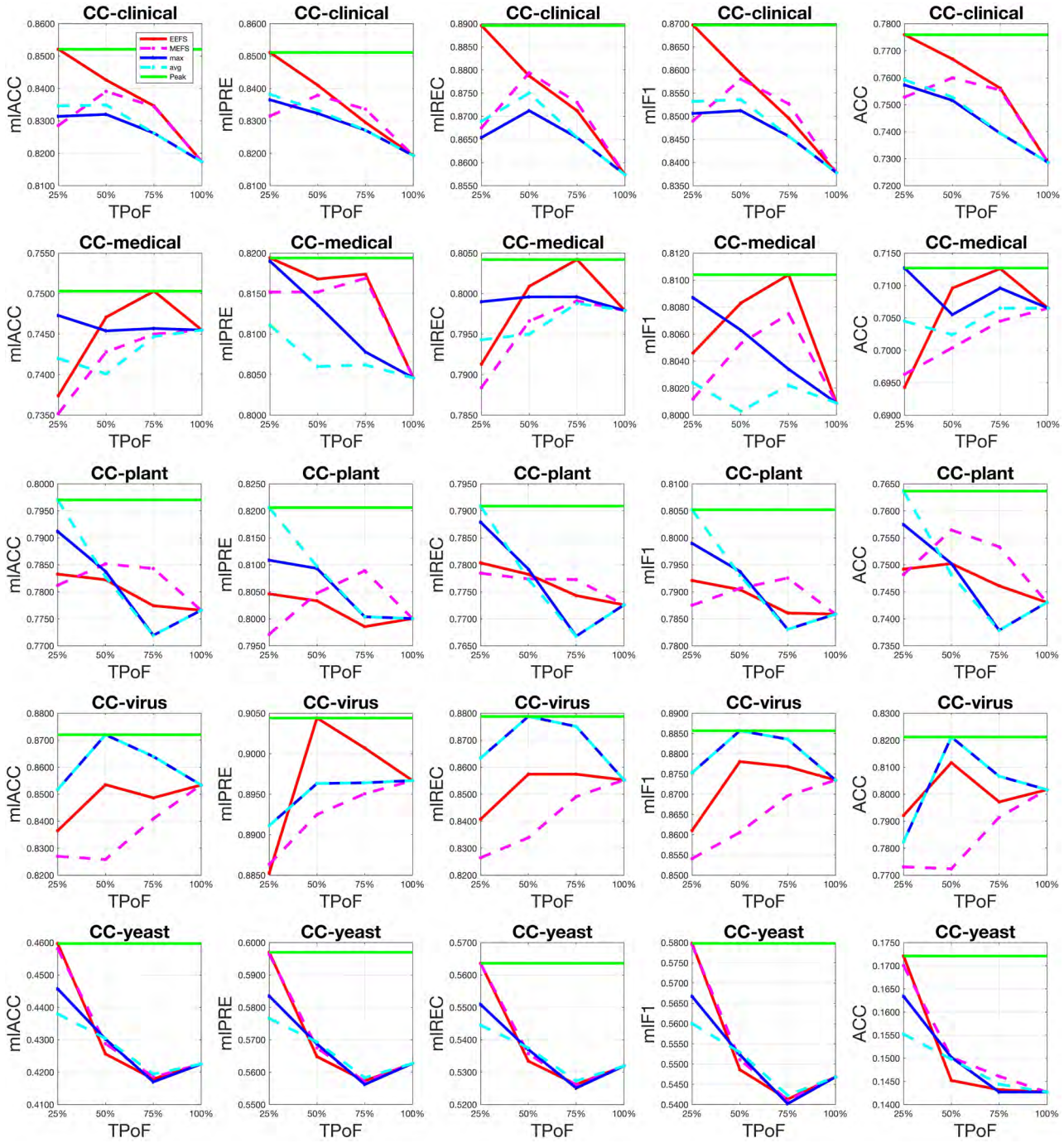


FIGURE 3. Performance of the four algorithms with BR classifier on five datasets (TPoF: Top percentage of feature ranking in descending order according to their importance).

V. RESULTS ANALYSIS

In this section, we will analyze the experimental results in detail. All results with 3 classifiers, 4 algorithms and 5 evaluation measures on five multi-label bioinformatics datasets are demonstrated in Table 1-3 and Figure 1-4. In Table 1-3, the optimal results means the selected best performance of each algorithms among the four top percentages (25%, 50%, 70%, 100%) of features with corresponding multi-label classifiers and evaluation measures. The bold-faced values represent the best performance among all the algorithms in Table 1-3. We use *benchmark* represent the cross validation classification results of corresponding multi-label classifiers and evaluation measures about the related datasets with all the features (no feature selection). As shown in Table 1-3, compared with *benchmark*, EEFS ranks 1st in 90.7% cases (BR: 76.0%, CC: 100%, MLkNN: 96.0%) and equally 1st

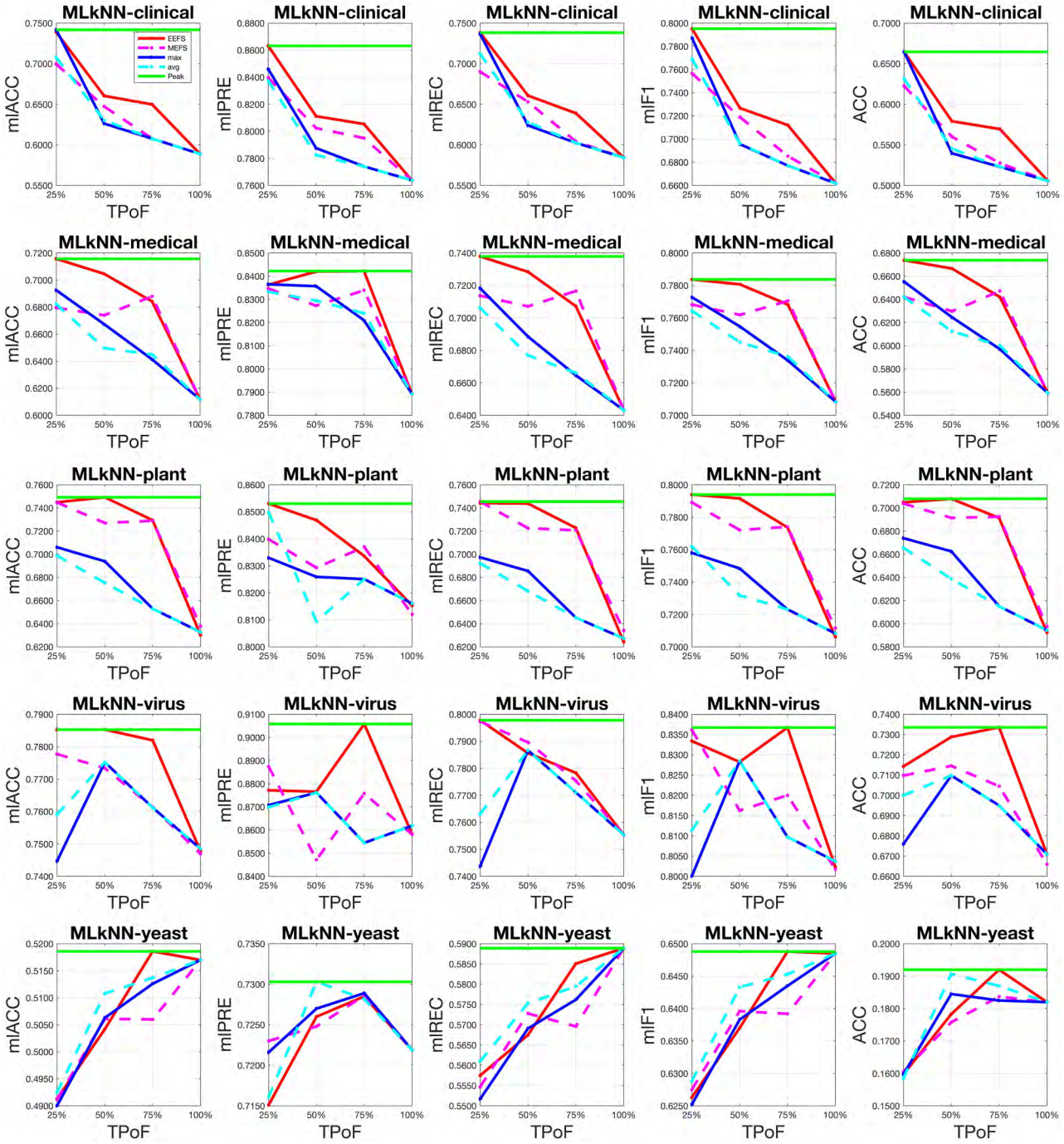


FIGURE 4. Performance of the four algorithms with BR classifier on five datasets (TPoF: Top percentage of feature ranking in descending order according to their importance).

in 9.3% cases (BR: 24.0%, CC: 0%, MLkNN: 4.0%) which demonstrates the multi-label feature selection effectiveness of EEFS. As shown in Table 1-3, and Figure 1, compared with other three algorithms, EEFS ranks 1st in 72.0% cases (BR: 72.0%, CC: 60.0%, MLkNN: 84.0%) and equally 1st in 2.7% cases (BR: 4.0%, CC: 0%, MLkNN: 4.0%). In Figure 1, each dataset under BR, CC and MLkNN classifiers connects all the evaluation measures with different color curves simultaneously. Different color curves represent different rank status and the number of color curves on the left half circle denotes the best performance of EEFS corresponding to each evaluation measure compared with other three algorithms. For example, under BR, we can see EEFS achieves 5 blue curves (ranking 1st) on *clinical* for five evaluation measures, 1 blue curve (ranking 1st) on *medical* for *mlREC*, and 4 yellow curves (ranking 3rd) on *medical* for other four evaluation measures. In Figure 2-4, the x-axes represent the top percentage of the feature importance ranking

TABLE 4. Characteristics of the experimental datasets.

Data set	$\|\mathcal{S}\|$	$dim(\mathcal{S})$	$L(\mathcal{S})$	$F(\mathcal{S})$	Domain
clinical	1566	232	10	numeric	text
medical	978	217	20	nominal	text
plant	969	224	12	numeric	biology
virus	206	185	6	numeric	biology
yeast	2417	103	14	numeric	biology

numbers and y-axes represent the performance values of different multi-label evaluation measures. The turning points of different color lines represent corresponding classification results of the selected feature subsets with different feature selection algorithms. The green lines (Peak) represent the best results in each evaluation measure. Based on the above experimental results, the following observations can be apparently made:

A. PERFORMANCE COMPARISON BETWEEN EEFS AND MEFS:

The optimal results of EEFS and MEFS shown in Table 1-3 demonstrate that our algorithm EEFS, which ranks 1st in 81.3% cases (BR: 72.0%, CC: 84.0%, MLkNN: 88.0%) and equally 1st in 8.0% cases (BR: 20.0%, CC: 0%, MLkNN: 4.0%), is absolutely better than MEFS with all the 3 classifiers on 3 multi-label bioinformatics datasets and 5 evaluation measures. As shown in Figure 2-4, when the size of feature subset is small which is top 25% features, EEFS ranks 1st in 80.0% cases (BR: 76.0%, CC: 84.0%, MLkNN: 80.0%), which indicates that EEFS can better evaluate the feature importance than MEFS. The superior performance of EEFS against MEFS clearly verifies the effectiveness of reducing the accumulated errors of data itself which could improve the performance of feature selection.

Furthermore, according to the algorithm computational complexity and mechanism, EEFS is higher computational efficiency than MEFS. For example, for training data with n instances, d features and q labels, to get the feature ranking, EEFS needs to train λ models with $\mu * n$ instances for each model and test d times to get each feature's importance for each model. MEFS needs to train $(d-1)$ models to get a feature ranking. In detail, the ith model of MEFS which is trained based on $(d-i+1)$ features with n instances needs test $(d-i+1)$ times to get each feature's importance. In experiments, MEFS is more time consuming than EEFS in getting the final feature ranking part and they are the same in other parts.

B. PERFORMANCE COMPARISON AMONG EEFS, MAX AND AVG:

As shown in Table 1-3, compared with *max* and *avg* across all evaluation measures and classifiers, EEFS ranks 1st in 73.3% cases (BR: 72.0%, CC: 60.0%, MLkNN: 80.0%). As shown in Figure 2-4, when the size of feature subset is small which is top 25% features, EEFS ranks 1st in 60.0% cases (BR: 64.0%, CC: 48.0%, MLkNN: 68.0%). These phenomenons, when top 25% features are selected, indicate that: 1) For BR and MLkNN, EEFS can better evaluate the feature importance than *max* and *avg*; 2) For CC, EEFS is not good as *max* and *avg*. When we analyze it in detail, we find the performance of EEFS on *plant* and *virus* impacts the results of feature selection. For *plant* and *virus* on top 25% features, EEFS ranks 1st in 70% cases with BR and 100% cases with CC, but it ranks 1st in 10% cases with CC. This phenomenon is because of the characteristics of CC classifier and *go* features of the protein datasets. CC employs label as new feature with original feature space to predict next label, but the labels structure do not match the structure of *go* features. Finally, it gets the poor performance.

All results indicate that, during the process of multi-label feature selection, EEFS can utilize: 1) The correlations between multiple labels and features; 2) The correlations within labels. In contrast, the other two algorithms, *max* and *avg*, are implemented by transforming sing-label methods to multi-label methods according to Eq. (1) and Eq. (2), respectively. Therefore, they can only utilize the relationship between single label and single feature which leads to their worse performance than EEFS.

VI. CONCLUSION

In this paper, we propose a novel algorithm named EEFS, i.e. *Ensemble Embedded Feature Selection*, which can deal with the multi-label feature selection problems in bioinformatics data. EEFS can provide relatively stable ranking of feature importance and reduce the negative effect from the change of training data by randomly selecting partial training examples and utilizing iteration to compute the prediction risk. As illustrated in the experimental results of most cases, the performance of EEFS is better than MEFS because of that it can reduce the accumulated errors of data itself by employing ensemble method. And it is better than other two filter multi-label feature selection algorithms, i.e., *max* and *avg* because of that it can utilize: 1) the correlations between multiple labels and features, 2) the correlations within labels.

Inspired by this work, we will further explore the mechanism of embedded multi-label feature selection methods for bioinformatics data and propose a more efficient algorithm in the future.

REFERENCES

[1] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," *Data Mining Knowl. Discovery Handbook*, vol. 7, pp. 667–685, Jul. 2010.

[2] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 681–687.

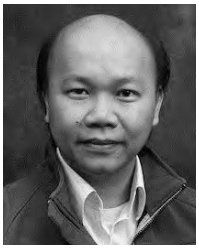
[3] G.-Z. Li, X. Wang, X. Hu, J.-M. Liu, and R.-W. Zhao, "Multilabel learning for protein subcellular location prediction," *IEEE Trans. Nanobiosci.*, vol. 11, no. 3, pp. 237–243, Sep. 2012.

[4] C. Xiang, X. Xuan, and K. C. Chou, "PLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC," *Genomics*, vol. 110, no. 1, pp. 50–58, Jan. 2017.

[5] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings Bioinf.*, vol. 6, no. 1, pp. 57–71, 2005.

[6] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "Frontiers of biomedical text mining: Current progress," *Briefings Bioinf.*, vol. 8, no. 5, pp. 358–375, Sep. 2007.
[7] Z. Wang, J. Shawe-Taylor, and A. Shah, "Semi-supervised feature learning from clinical text," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2010, pp. 462–466.
[8] H. Nassif, F. Cunha, I. C. Moreira, R. Cruz-Correia, E. Sousa, D. Page, E. Burnside, and I. Dutra, "Extracting BI-RADS features from Portuguese clinical texts," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Philadelphia, PA, USA, Oct. 2012, pp. 1–4.
[9] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang, "Multi-label sparse coding for automatic image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1643–1650.
[10] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, "A MFoM learning approach to robust multiclass multi-label text categorization," in *Proc. 21st Int. Conf. Mach. Learn.*, Jul. 2004, p. 42.
[11] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Salvador, Brazil, Aug. 2005, pp. 258–265.
[12] S. Yu, K. Yu, V. Tresp, and H.-P. Kriegel, "Multi-output regularized feature projection," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 12, pp. 1600–1613, Dec. 2006.
[13] S. Ji, L. Tang, S. Yu, and J. Ye, "Extracting shared subspace for multi-label classification," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2008, pp. 381–389.
[14] C. H. Park and M. Lee, "On applying linear discriminant analysis for multi-labeled problems," *Pattern Recognit. Lett.*, vol. 29, no. 7, pp. 878–887, 2008.
[15] K. Daniilidis, P. Maragos, and N. Paragios, Eds., *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 6316. Heraklion, Greece: Springer, Sep. 2010.
[16] X. Shu, H. Xu, and L. Tao, "A least squares formulation of multi-label linear discriminant analysis," *Neurocomputing*, vol. 156, pp. 221–230, May 2015.
[17] Y. Zhang and Z.-H. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Trans. Knowl. Discovery Data*, vol. 4, no. 3, p. 14, Oct. 2010.
[18] D. Zhu and B. Carterette, "Improving health records search using multiple query expansion collections," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Oct. 2012, pp. 1–7.
[19] L. C. Molina, L. Belanche, and A. Nebot, "Feature selection algorithms: A survey and experimental evaluation," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2002, pp. 306–313.
[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[21] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
[22] Y. Guo, F. Chung, and G. Li, "An ensemble embedded feature selection method for multi-label clinical text classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 823–826.
[23] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. 14th Int. Conf. Mach. Learn.*, Jul. 1997, pp. 412–420.
[24] L. N. de Barros, M. Finger, A. T. R. Pozo, G. A. G. Lugo, and M. A. Castilho, Eds., *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 7589. Curitiba, Brazil: Springer, Oct. 2012.
[25] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, nos. 1–2, pp. 23–69, Oct. 2003.
[26] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 6, pp. 1437–1447, Nov. 2003.
[27] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[28] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, 1997.
[29] M. M. Kabir, M. M. Islam, and K. Murase, "A new wrapper feature selection approach using neural network," *Neurocomputing*, vol. 73, nos. 16–18, pp. 3273–3283, Oct. 2010.
[30] H. Shao, G. Li, G. Liu, and Y. Wang, "Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine," *Sci. China Inf. Sci.*, vol. 56, no. 5, pp. 1–13, May 2013.
[31] M. You, J. Liu, G.-Z. Li, and Y. Chen, "Embedded feature selection for multi-label classification of music emotions," *Int. J. Comput. Intell. Syst.*, vol. 5, no. 4, pp. 668–678, Aug. 2012.
[32] S. Wang, J. Tang, and H. Liu, "Embedded unsupervised feature selection," in *Proc. 29th AAAI Conf. Artif. Intell.*, Feb. 2015, pp. 470–476.
[33] S. Maldonado and J. López, "An embedded feature selection approach for support vector classification via second-order cone programming," *Intell. Data Anal.*, vol. 19, no. 6, pp. 1259–1273, Jan. 2015.
[34] J. P. Pestian, C. Brew, P. Matykiewicz, D. Hovermale, N. Johnson, K. B. Cohen, and W. Duch, "A shared task involving multi-label classification of clinical free text," in *Proc. Workshop BioNLP Biol., Transl., Clin. Lang. Process.*, Jun. 2007, pp. 97–104.
[35] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," in *Proc. ACM Symp. Appl. Comput.*, Apr. 2006, pp. 235–239.
[36] R.-W. Zhao, G.-Z. Li, J.-M. Liu, and X. Wang, "Clinical multi-label free text classification by exploiting disease label relation," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Dec. 2013, pp. 311–315.
[37] M. Moisio, *A Guide to Health Insurance Billing*. Clifton Park, NY, USA: Thomson Delmar Learning, 2000.
[38] K.-C. Chou and H.-B. Shen, "Plant-mPLoc: A top-down strategy to augment the power for predicting plant protein subcellular localization," *PloS One*, vol. 5, no. 6, 2010, Art. no. e11335.
[39] X. Cheng, X. Xiao, and K. C. Chou, "PLoc-mPlant: Predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC," *Gene*, vol. 13, no. 9, pp. 1722–1727, 2017.
[40] H.-B. Shen and K.-C. Chou, "Virus-mPLoc: A fusion classifier for viral protein subcellular location prediction by incorporating multiple sites," *J. Biomolecular Struct. Dyn.*, vol. 28, no. 2, pp. 175–186, Oct. 2010.
[41] X. Xiao, Z.-C. Wu, and K.-C. Chou, "ILoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites," *J. Theor. Biol.*, vol. 284, no. 1, pp. 42–51, Sep. 2011.
[42] K.-C. Chou and H.-B. Shen, "Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms," *Development*, vol. 109, no. 10, p. 1091, 2010.
[43] X. Cheng, X. Xiao, and K.-C. Chou, "PLoc-mVirus: Predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC," *Gene*, vol. 628, pp. 315–321, Sep. 2017.
[44] X. Cheng, X. Xiao, and K.-C. Chou, "PLoc-mHum: Predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information," *Bioinformatics*, vol. 34, no. 9, pp. 1448–1456, 2017.
[45] X. Cheng, X. Xiao, and K.-C. Chou, "Erratum to "PLoc-mVirus: Predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC" [Gene 628 (2017) 315-321]," *Gene*, vol. 644, p. 156, Feb. 2018.
[46] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.
[47] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *J. Mach. Learn.*, vol. 85, no. 3, pp. 333–359, Dec. 2011.
[48] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.
[49] X. Wang and G.-Z. Li, "Multilabel learning via random label selection for protein subcellular multilocations prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 2, pp. 436–446, Apr. 2013.
[50] X. Wang, W. Zhang, Q. Zhang, and G.-Z. Li, "MultiP-SChlo: Multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier," *Bioinformatics*, vol. 31, no. 16, pp. 2639–2645, 2015.
[51] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[52] M.-L. Zhang and L. Wu, "LIFT: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.

YUMENG GUO received the bachelor's degree in automation from Tongji University, Shanghai, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Control Science and Engineering, Tongji University, and the joint Ph.D. degree with the Department of Computing, The Hong Kong Polytechnic University. His current research interests include data mining, machine learning, multi-label learning, and bioinformatics.



FU-LAI CHUNG received the B.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, and the M.Phil. and Ph.D. degrees from The Chinese University of Hong Kong, Hong Kong. He is currently an Associate Professor with the Department of Computing, The Hong Kong Polytechnic University. He has published over 200 refereed papers in various international journals and conferences, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON FUZZY SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition*, *Neural Networks*, AAAI, IJCAI, SIGIR, and CIKM. His current research interests include deep learning, transfer learning, adversarial learning, graph mining, time series mining, recommendation systems, and computational creativity. He also serves on program committee of top international conferences, including the IEEE ICDM, AAAI, IJCAI, and ICPR.

GUOZHENG LI received the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, in 2004. He is currently the Chief Technical Officer with Shanghai Taikuntang Traditional Chinese Medicine Hospital. He is also a Principle Investigator of several projects under grants of the Natural Science Foundation of China. He has published over 100 refereed papers in journals and conferences. His research interests include pattern recognition and bio-medical data mining. He is also an Executive Committee Member of the CAAI Machine Learning Society.

LEI ZHANG received the M.D. degree from the Shandong University of Traditional Chinese Medicine, in 2011. From 2011 to 2013, he was a Postdoctoral Fellow with the School of Computer and Information Technology, Beijing Jiaotong University. Since 2013, he has been a Research Assistant with the Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences. His main research interest includes clinical data mining in traditional Chinese medicine.