Pattern Recognition 67 (2017) 410–423

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Granular multi-label feature selection based on mutual information

Feng Li[a,b,c], Duoqian Miao[a,b,*], Witold Pedrycz[c]

[a] Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China,
[b] Key Laboratory of Embedded System and Service Computing, Ministry of Education of China, Tongji University, Shanghai, 201804, China
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, T6G 1H7, Canada.

ARTICLE INFO

Article history:
Received 31 May 2016
Revised 17 February 2017
Accepted 18 February 2017
Available online 27 February 2017

Keywords:
Granular computing
Feature selection
Multi-label learning
Mutual information

ABSTRACT

Like the traditional machine learning, the multi-label learning is faced with the curse of dimensionality. Some feature selection algorithms have been proposed for multi-label learning, which either convert the multi-label feature selection problem into numerous single-label feature selection problems, or directly select features from the multi-label data set. However, the former omit the label dependency, or produce too many new labels leading to learning with significant difficulties; the latter, taking the global label dependency into consideration, usually select a few redundant or irrelevant features, because actually not all labels depend on each other, which may confuse the algorithm and degrade its classification performance. To select a more relevant and compact feature subset as well as explore the label dependency, a granular feature selection method for multi-label learning is proposed with a maximal correlation minimal redundancy criterion based on mutual information. The maximal correlation minimal redundancy criterion makes sure that the selected feature subset contains the most class-discriminative information, while in the meantime exhibits the least intra-redundancy. Granulation can help explore the label dependency. We study the relation of the label granularity and the performance on four data sets, and compare the proposed method with other three multi-label feature selection methods. The experimental results demonstrate that the proposed method can select compact and specific feature subsets, improve the classification performance and performs better than other three methods on the widely-used multi-label learning evaluation criteria.

1. Introduction

Traditional machine learning algorithms usually aim to find the most relevant class label l of the unseen instance, which are called single-label learning algorithms. Unlike the single-label learning, given a predefined label space L, the task of multi-label learning algorithms is to predict a set of relevant labels Y of the unseen instance, where $Y \subseteq L$ and $|Y| \geq 1$ [1–4].

Multi-label instances exist in various real-world domains. For example, in the image domain [5], an image may exhibit multiple semantic classes, such as *city* and *sea*. In the text domain [6], a document may belong to several predefined topics simultaneously, such as, *society* and *entertainment*. In the biology domain [7], a gene could have a set of functions, such as *metabolism* and *transcription*. What's more, the labels usually exhibit some correlations, for example, an image containing *beach* has a greater possibility to contain *sea* simultaneously than to contain *grassland*.

As in the traditional machine learning algorithm, the classification performance of multi-label learning algorithm greatly depends on the quality of available features, and the algorithm is often faced with the curse of dimensionality [8–10]. The data sets used for learning usually contain a large number of features, some of which are irrelevant or redundant for classification purpose. The irrelevant or redundant features do not only increase the algorithm complexity, but also degrade the classification performance [11]. Feature selection is a process of reducing dimensionality of the feature space, which aims to select a subset of the most relevant features without losing class-discriminative information, and remove irrelevant and redundant features [12].

Feature selection is an important problem for pattern classification systems, and there are a large number of approaches to feature selection for machine learning, however most of them are dedicated for single-label learning, for example, Peng [13] proposed a min-redundancy and max-relevance criterion (mRMR) based on mutual information for feature selection in single-label learning. Recently, as the multi-label learning has attracted more attention, some scholars stressed the relevance of feature dimension reduction problem in multi-label learning. One commonly en-

* Corresponding author.
E-mail addresses: dqmiao@tongji.edu.cn, miaoduoqian@163.com (D. Miao).

http://dx.doi.org/10.1016/j.patcog.2017.02.025
0031-3203/

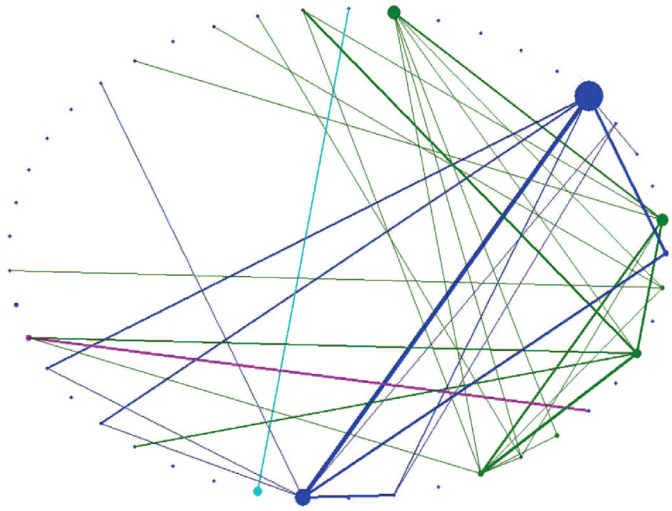

Fig. 1. Label correlation.

countered way dealing with the multi-label learning is to transform the multi-label problem into numerous single-label problems [14,15], and then the relevant features for each transformed new single label can be easily selected with the aid of any traditional single-label methods [16–18]. However, this way either ignores the correlation among labels which may cover very important information contained in the multi-label data, or creates too many new labels, resulting in the difficulty of the ensuing learning. To solve the difficulty arising in this way, some feature selection methods have been proposed to deal with multi-label data directly by evaluating the feature in term of dependency with all labels [19–22]. However, the proposed methods directly handling multi-label learning usually select features by considering the entire set of labels at a time, while there does not always exist a strong correlation between any pair of labels, especially in data with a large collection of labels. Thus the selected feature may be very useful for a single label, while insignificant or even harmful for another label with limited dependency, which may make the selected feature subset too complicated and degrade the performance. Furthermore, the complexity of the algorithm for feature selection and classification may be strongly impacted by the number of labels.

For example, Fig. 1 displays the statistical information about the labels of the *Medical* data set. *Medical* is a multi-label data set about medical diagnosis [23], which has been used in Computational Medicine Centers 2007 Medical Natural Language Processing Challenge. It includes a free-text summary of patient clinical history, impression and their prognosis, labeled with ICD-9-CM codes. There are 45 possible labels in *Medical* data set, and each label is shown as a node of the graph. The node size is in proportion to the number of instances belonging to the corresponding label. An edge of the graph indicates that there exists more than a single instance simultaneously associated with the two labels linked by the edge, and the width of the edge relates to the number of instances. For a concise structure illustration, any pair of labels simultaneously associated with just one instance is not shown in this figure. As seen, there are four main groups of labels marked with different colors, and different groups almost do not have any links in Fig. 1. The rest of the remaining nodes are independent.

Therefore, to prudently explore the label correlation, while select a more relevant and compact feature subset, we propose a granular feature selection method for multi-label learning with a maximal correlation minimal redundancy criterion based on mutual information. Granular computing(GrC) is about processing complex information entities called information granules, which arise in the process of data abstraction and derivation of knowledge from data [24–27]. Information granule is the generic construct of granular computing. Firstly, the labels can be abstracted into several information granules according to their correlations. The relevant labels are placed into the same information granule by considering the dependency, which doesn't exhibit a significant degradation on the classification performance, for labels in the same information granule have the largest dependency, while labels in different information granules have very little dependency. Then, for an information granule, an immediate method is to select features having the maximal dependency on this information granule, while having minimal intra-redundancy. However, there are numerous labels in an information granule, so it is not easy to take all label combinations into computing the dependency directly. An approximated way is to choose the feature having the largest correlation with each label in the information granule while minimal correlation with already selected features, called maximal correlation minimal redundancy criterion. Maximal correlation criterion guarantees that we pick out the most related features and omit irrelevant features, while minimal redundancy criterion makes sure redundant features are not selected. Finally, a specific feature subset is separately selected for each information granule of labels by this way, which is more relevant and compact. The method does not only consider the correlation among labels, but also selects more specific feature subsets for labels to improve the classification performance. The main contributions of this study include:

- We consider a granulation method to granulate labels of the label space into several information granules to explore the local dependency among labels and choose a specific and compact feature subset for each information granule of labels, on which there are few researches.
- We propose a feature selection method for multi-label learning based on mRMR [13], which does not only maximize the dependency between the feature subset and the target information granule, but also minimizes the redundancy in the feature subset. We also prove that the maximal dependency of the feature and the target information granule can be approximated by maximizing the sum of the individual correlation of the feature and each label in the target information granule.
- The comprehensive experiments are involved to verify the effectiveness of the label granularity on the size of the optimal feature subset and the classification performance on four multi-label data sets. Furthermore, we compare the proposed method with other three feature selection methods. The results show that the proposed method can choose compact and specific feature subsets, and achieve better performances in comparison with the performance of the existing methods.

2. Related studies

Feature dimensionality reduction mechanisms can be mainly divided into feature extraction and feature selection. Recently, feature extraction technique, including principle component analysis (PCA) [19] and linear discriminant analysis (LDA) [28], has been considered to reduce feature dimensions in multi-label classification. For example, Zhang and Zhou [19] proposed a dimensionality reduction method for multi-label naive Bayes classification (MLNB). It firstly eliminated irrelevant and redundant features by PCA. Then, the appropriate subset of features was selected for classification using a genetic algorithm.

Among proposed multi-label algorithms, one popular way is to transform the multi-label learning task into several single-label learning tasks, such as binary relevance(BR) [29], label powerset(LP) [30], and Pruned Problem transformation (PPT) [31], which is called *Problem transformation method* [1], and then the resulting

problem can be easily solved by single-label learning algorithms [32–34]. Similarly, some feature selection methods for multi-label learning transform the multi-label data into single-label data, then select the relevant features for each label. Binary relevance(BR) transformation is used to create a binary single-label data set for each label in [35], and then feature subset is selected for each single-label data set with relief and information gain measures. However, this method may yield poor classification performance because it handles the labels independently, leading to information losses about correlation among labels. The label powerset(LP) is applied to transform the music multi-label data set *Emotions* into a single-label data set in [15]. The most relevant features are then selected in accordance with the dependency between feature and label, which is measured by χ^2 statistics. Although the LP considers the correlation among labels, it may create too many new labels, and cause the over-fitting and imbalance problem [36].

The problem transformation method usually does not take the label correlation into account, or create numerous new labels, which may degrade the classification performance. Therefore, some feature selection methods are put forward to select features on multi-label data set directly. In [21], Lee put forward a feature selection method for multi-label classification based on multivariate mutual information (MUMI). An effective feature subset is selected by maximizing the dependency between selected features and labels, and label interaction is considered without resorting to problem transformation. Lin [22] proposed a multi-label learning feature selection based on max-dependency and min-redundancy (MDMR). The method chooses feature according to the dependency between the candidate feature and all labels and the conditional redundancy between the candidate feature and the selected features simultaneously. A fast multi-label feature selection based on information-theoretic feature ranking algorithm was proposed in [37], where features were ranked according to their importance, and then the top-ranked features were chosen. Almost all methods directly selecting features from multi-label data set consider the global dependency among labels, but not all labels greatly depend on each other. The selected features when taking all labels into consideration may have very important discriminative information for one label, while may not exhibit any discriminative value for another label which has little dependency on the label. Meanwhile, it is not easy work to compute the dependency between each pair of labels, especially when data sets have a large number of labels.

3. Preliminaries

In this section, we briefly recall the main concepts, which will be used in further discussion.

3.1. Mutual information

The Shannon's entropy [38,39] is a measure of uncertainty of random variables. The entropy of a random variable X is defined as:

$$H(X) = -\sum_{x\in X} p(x)\log p(x), \tag{1}$$

where $p(x)$ is the probability of x. The joint entropy between two variables X and Y is defined as follows:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y), \tag{2}$$

where the $p(x, y)$ is the joint probability of x and y.

When the variable X is known while Y is not, the conditional entropy can measure the remaining uncertainty, which is defined as:

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x), \tag{3}$$

where $p(y|x)$ is the conditional probability. Here, the conditional entropy can be replaced by entropy and joint entropy as follows:

$$H(Y|X) = H(X,Y) - H(X). \tag{4}$$

Mutual information is one of the widely used measures of dependency of variables. More specifically, it quantifies the obtained information of one random variable, through the other random variable. The mutual information between two random variables is defined as follows:

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{5}$$

If the value of mutual information between two random variables is high, this means that the two variables are highly dependent.

Lemma 1. *Mutual information can be represented by the corresponding individual entropies as* [40]:

$$\begin{aligned} MI(X;Y) &= H(Y) - H(Y|X) = H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned} \tag{6}$$

Proof.

$$\begin{aligned} MI(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x)p(y) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) \\ &\quad - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) - \sum_{x\in X} p(x)\log p(x) \\ &\quad - \sum_{y\in Y} p(y)\log p(y) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned}$$

□

Assuming that we already have variable Z, the mutual information between X and Y given the value of variable Z, called conditional mutual information, is defined as follows [40]:

$$MI(X;Y|Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)log\frac{p(x,y|z)}{p(x|z)p(y|z)}. \tag{7}$$

Lemma 2. *There exists a relationship between conditional mutual information and mutual information coming in the form* [40]:

$$MI(X;Y|Z) = MI(X;Y,Z) - MI(X;Z). \tag{8}$$

Proof.

$$\begin{aligned} MI(X;Y|Z) &= \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)log\frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \\ &= \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)log\frac{p(x)p(z)p(x,y,z)}{p(x)p(x,z)p(y,z)} \\ &= \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)log\frac{p(x,(y,z))}{p(x)p(y,z)} \\ &\quad - \sum_{x\in X}\sum_{z\in Z} p(x,z)log\frac{p(x,z)}{p(x)p(z)} \\ &= MI(X;Y,Z) - MI(X;Z). \end{aligned}$$

□

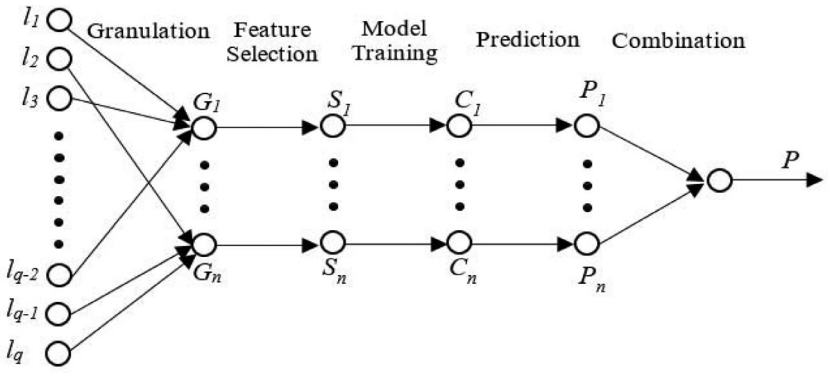

Fig. 2. The structure of the proposed algorithm.

4. Proposed algorithm

In multi-label learning, let $F = R^b$ stand for the b-dimensional input feature space, and $L = \{l_1, l_2, \cdots, l_q\}$ denote the predefined label space. Each instance $X_i \in F$ can be represented as a b-dimensional feature vector $X_i = \{x_1^i, x_2^i, \cdots, x_b^i\}$, which is associated with a set of labels $Y_i \subseteq L$. By learning from a training multi-label data set $Tn = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_a, Y_a)\}$, the multi-label learning algorithm assigns the relevant labels to the unseen instance, while the algorithm performance is significantly related with the input features.

To improve the algorithm performance, a granular feature selection method is proposed to select the most dependent features. The proposed feature selection method is independent of classifier algorithm. Firstly, the labels in label space L are granulated into several information granules G based on their similarity information contained in the training data set Tn to select more specific and relevant features as discussed in Section 4.1. For each label information granule, we can get a feature subset having the maximal dependency on the labels with the maximal correlation minimal redundancy criterion, introduced in Section 4.2, then train a classifier model for each pair of feature subset and label information granule. For an unseen instance, the output is the combination of the prediction achieved by each classifier model. The overall structure of the algorithm is shown in Fig. 2.

4.1. Granulation

In granular computing, information granules are collections of entities that usually originate at the numeric level and are arranged together due to their similarity, functional or temporal adjacency, indistinguishability, coherency or the like. In this paper, the labels are arranged into information granules G according to their similarity. The labels in one information granule are relevant to each other, while labels in different information granules are irrelevant, so that the label correlation is explored as much as possible. The size of G is smaller than the size of the label space L. Imbalance problem is an inherent property in multi-label data, and it is very likely that one information granule contains much more labels than another one. To avoid a lack of balance problem among labels, the size of each information granule is limited to evenly arrange labels into information granules. For balanced clustering, a balanced k-means method has been proposed in [41]. K-means clustering is a popular technique. Here the k-means method is used to realize granulation refer to Algorithm 1. Then a feature subset S for each label information granule G is selected using the feature selection technique.

Algorithm 1 Label granulation.

Input: Number of information granules n; Label space L; Label data W; iterations it;
Output: n label information granules

```
for i = 1 to n do
  //Initialize information granules G_i and information granule centers g_i
  G_i ← φ
  g_i ← randomly selected member of L
end for

while it > 0 do
  for each l_j ∈ L do
    for k = 1 to n do
      d_kj ← distance(l_j, g_k, W)
    end for
  end for
  No ← |L|
  while No > 0 do
    Find arg min_{1≤k≤n, l_j∈L} d_kj
    if |G_k| ≥ ⌈|L|/n⌉ then
      d_kj ← ∞
    else
      Add l_j to the information granule G_k
      d_*j ← ∞
      No ← No − 1
    end if
  end while
  Calculate centers of information granule g'_k
  if all g'_k == g_k then
    break;
  else
    G_i ← φ
    g_k ← g'_k
  end if
  it ← it − 1
end while
Output the label information granules G_1, G_2, ···, G_n;
```

4.2. Maximal correlation minimal redundancy criterion

The crux of feature selection is to select a feature subset S consisting of m features coming from the input feature set $F(m < b)$, which jointly have the maximal dependency on the target label information granule G. Based on mutual information, the dependency is measured by $MI(S; G)$. To ensure the maximal dependency between S and G, every selected feature should exhibit the largest

contribution to $MI(S; G)$. Suppose a subset S_{m-1} with $m-1$ features has been picked, and the label information granule G contains r labels. The mth feature is the one with the maximal mutual information with label information granule G given the feature subset S_{m-1}, which is quantified by conditional mutual information expressed in the form:

$$\max MI(G; x_m|S_{m-1}) = \max[MI(G; S_{m-1} \cup \{x_m\}) - MI(G; S_{m-1})], \tag{9}$$

where the $MI(G; S_m)$ is calculated as follows:

$$\begin{aligned} MI(S_m; G) &= \sum\sum p(S_m, G) log \frac{p(S_m, G)}{p(S_m)p(G)} \\ &= \sum\sum\sum p(S_{m-1}, x_m, G) log \frac{p(S_{m-1}, x_m, G)}{p(S_{m-1}, x_m)p(G)} \\ &= \sum\cdots\sum p(x_1, \cdots x_m, l_1, \cdots l_r) \\ &\quad \times log \frac{p(x_1, \cdots x_m, l_1, \cdots l_r)}{p(x_1, \cdots x_m,)p(l_1, \cdots l_r)}. \end{aligned} \tag{10}$$

It can be seen that the computing cost increases exponentially as the number of features and labels increase. Here comes the difficultly when directly implementing the maximal dependency. A min-redundancy and max-relevance criterion(mRMR) is proposed in [13] to solve this problem. Based on mRMR, we consider a feature selection method for multi-label learning.

To maintain as much class-discriminative power about the label information granule G as possible, the selected feature subset S should be maximally related to G, specifically to all labels in G, which is called maximal correlation criterion, a approximation of the max dependency. The correlation is expressed in terms of mutual information. The selected S satisfies the following condition with the mean value of mutual information between S and G:

$$\max D(S, G), D = \frac{1}{|S|}\sum_{x_i \in S} MI(x_i; G). \tag{11}$$

From the above expression, one can guarantee the maximal correlation criterion stating that each new selected feature x_i exhibits the maximal correlation to G, so the condition can be converted into the following one:

$$\max D(x_i, G), D = MI(x_i; G). \tag{12}$$

Unlike traditional classification problem having only a single label, there are usually multiple labels in the information granulation G. It is hard to implement the above criterion, especially when the size of G is large. Intuitively, the maximal correlation criterion between feature x_i and information granule G shown in Eq. (12) is approximately equal to the criterion stating that feature x_i has the maximal correlation with each label $l_j \in G$. This fact has been proved in the appendix. Thus, the condition can be rewritten as:

$$\max\ C(x_i, G), C = \frac{1}{|G|}\sum_{l_j \in G} MI(x_i; l_j). \tag{13}$$

According to the maximal correlation criterion, it is likely that the new selected feature may exhibit a significant dependency on some features of the selected feature subset, thus increasing intra-redundancy in S. When two features show significant dependency, the respective class-discriminative power would not increase much if one of them has been added. Therefore the selected feature should have the largest correlation to the label information granule G, while it should exhibit the minimal dependency on the selected feature subset S. Therefore, the following minimal redundancy condition can be added when selecting a new feature.

$$\min R(x_i, S), R = \frac{1}{|S|}\sum_{x_j \in S} MI(x_i; x_j). \tag{14}$$

Table 1
Multi-label data sets used in the experiments.

Name	Instance	Feature	Label	Cardinality	Density
Enron	1702	1001	53	3.378	0.064
Medical	978	1449	45	1.245	0.028
Emotions	593	72	6	1.869	0.311
Genbase	662	1185	27	1.252	0.046

Combining the two constraints specified above, the operator $\Gamma(C, R)$ is defined to combine C and R, and consider the following simpler form to optimize C and R simultaneously:

$$\max \Gamma(C, R), \Gamma = C - R. \tag{15}$$

Given a feature subset S and the label information granule G, the task of feature selection is to find the ith feature from the set $F - S$ when maximizing the $\Gamma(\cdot)$. The $\Gamma(\cdot)$ can be formed by using (13) and (14) as:

$$\Gamma(C, R) = \frac{1}{|G|}\sum_{l_j \in G} MI(x_i; l_j) - \frac{1}{|S|}\sum_{x_j \in S} MI(x_i; x_j). \tag{16}$$

The method is realized in the form of Algorithm 2.

Algorithm 2 Granular feature selection for multi-label learning.

Input:
Feature set F; The size of feature subset m;
Label information granule G;
Output: Feature subset S
// Initialize the feature subset
$S = \phi$;
for $i = 1$ to b **do**
Compute the mutual information $MI(x_i; l_j)$ between feature x_i and label l_j in G;
Compute the mutual information $MI(x_i; x_j)$ between feature x_i and each feature x_j;
end for
repeat
Find the feature $x_i \in F$ satisfying Eq. (15)
$S = S \cup \{x_i\}$;
$F = F \setminus \{x_i\}$;
until $|S| == m$
Output the selected feature subset S for label information granule G;

Lemma 3. *The label information granule G with a coarser granularity has a larger size of the best feature subset S.*

Proof. If a label $l_j \in L$ is added to G represented by G', then G' is coarser than G. Suppose S, S_j and S' are the best feature subsets for G, l_j and G' respectively, $S' = S \cup S_j$, therefore $|S'| \geq |S|$. $|S'| = |S|$ if and only if $S_j \subseteq S$. □

5. Experimental studies

5.1. Data sets and experimental setting

To evaluate the performance of the proposed method, we experiment with four multi-label data sets covering different domains, as shown in Table 1[42]. In the table, the 'name' is the name of the data set and the 'instance', the 'feature' and the 'label' stand for the instance number, the feature number, and the label number of the data set, separately. The 'cardinality' is the *label cardinality* which is the average number of labels associated with each instance. The 'density' is the *label density* which is the average number of labels

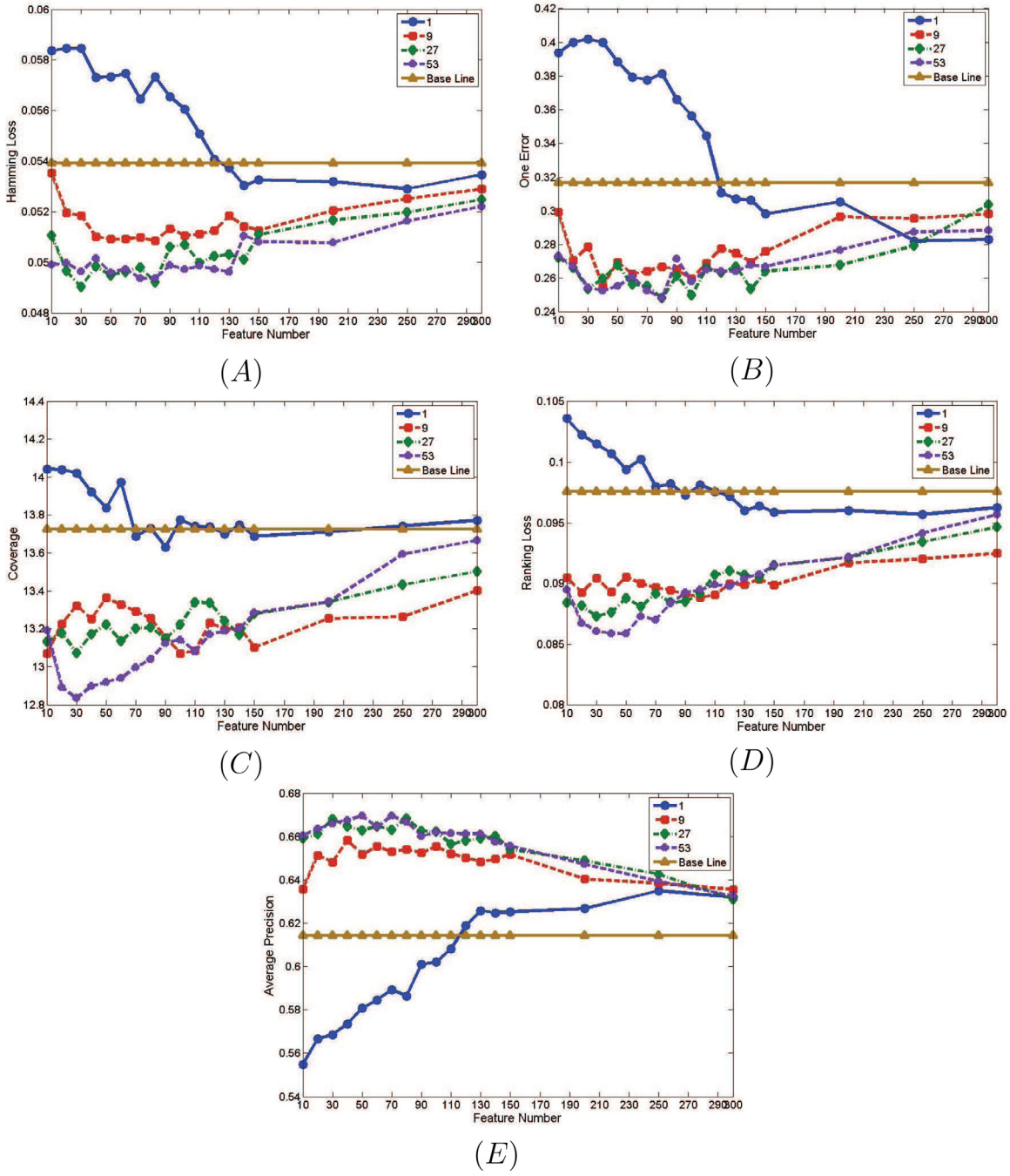

Fig. 3. Classification performance on *Enron* according to the number of selected features using the proposed method with different numbers (1,9,27,53) of information granules: (*A*) Hamming loss; (*B*) One error; (*C*) Coverage; (*D*) Ranking loss; (*E*) Average precision.

of instances data set divided by the number of labels in the data set.

Enron data set comes from the UC Berkeley Enron Email Analysis Project [43], and contains 1702 multi-labeled Enron e-mails, where the e-mails messages are made public from the Enron corporation. There are 53 possible tags in the data, such as Empty Message, Company Business, and Purely Personal.

Medical [23] data set contains 978 instances, and each instance is a radiology text report consisting of the medical history and symptom and associated with a subset of 45 ICD-9-CM labels.

Emotions [15] data set contains 593 instances, each of which is a song and represented by 8 rhythmic features and 64 timbre features. The instances are labeled with 6 possible emotions.

Genbase [44] data set consists of 662 proteins, and each protein chain is represented using a 1185 motif sequence vocabulary. The labels are 27 protein function families, such as a class of receptors, a class of oxydoreductases, and a class of transferases.

After selecting features, a classifier is used to produce the output. Zhang [45] combined the k-nearest neighbors algorithm and Bayesian inference to propose a multi-label lazy learning approach, i.e. multi-label k-nearest neighbors algorithm(ML-KNN). ML-KNN is a well known multi-label classification scheme for its efficiency. To produce the output, the ML-KNN method is chosen to complete classification. As recommended in [45], the number of nearest neighbors is set to 10, and the smoothing factor equals to 1. To obtain stable results, *ten-fold cross-validations* evaluation is used in

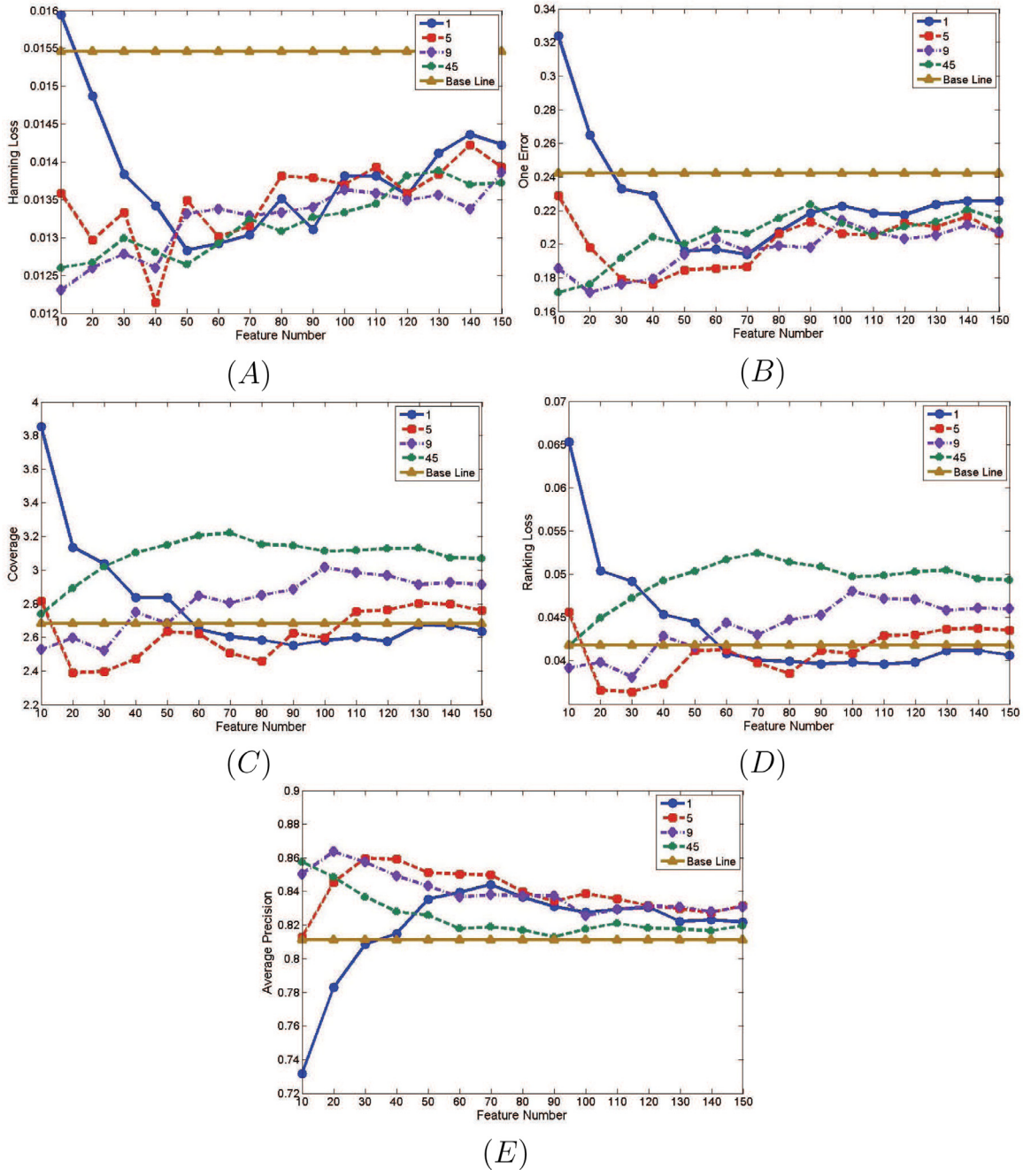

Fig. 4. Classification performance on *Medical* according to the number of selected features using the proposed method with different numbers (1,5,9,45) of information granules: (*A*) Hamming loss; (*B*) One error; (*C*) Coverage;(*D*) Ranking loss; (*E*) Average precision.

the experiment, and the average classification performance is reported.

5.2. Evaluation criteria

Different from the single-label learning algorithm, the performance evaluation for the multi-label learning algorithm is more complicated. The following five widely used multi-label evaluation criteria are used to quantify the performance in this paper, namely *Hamming loss, coverage, one-error, ranking loss* and *average precision* [46]. Given a test data set $Ts = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_t, Y_t)\}$, the multi-label learning algorithm predicts a relevant label set $Y_i^{'}$ for each unseen instance X_i, and then the used evaluation criteria are defined as:

(1) *Hamming Loss*: it evaluates the average difference between the predicted label set $Y_i^{'}$ and the ground-truth label set Y_i, i.e, a predicted label is not actually belonging to the instance or an unpredicted label is actually belonging to the instance. The performance of *Hamming Loss* is defined as:

$$\gamma_{HammingLoss} = \frac{1}{tq}\sum_{i=1}^{t}|Y_i^{'} \bigtriangleup Y_i|, \tag{17}$$

where $\triangle$ represents the difference between two sets, and $|\cdot|$ stands for the cardinality of a set. The smaller the value of Hamming loss, the better the performance. The algorithm achieves the best performance, when the Hamming loss attains 0.

(2) *Coverage*: the measure evaluates how far we need, on average, to go down the ranked list of labels in order to cover all the

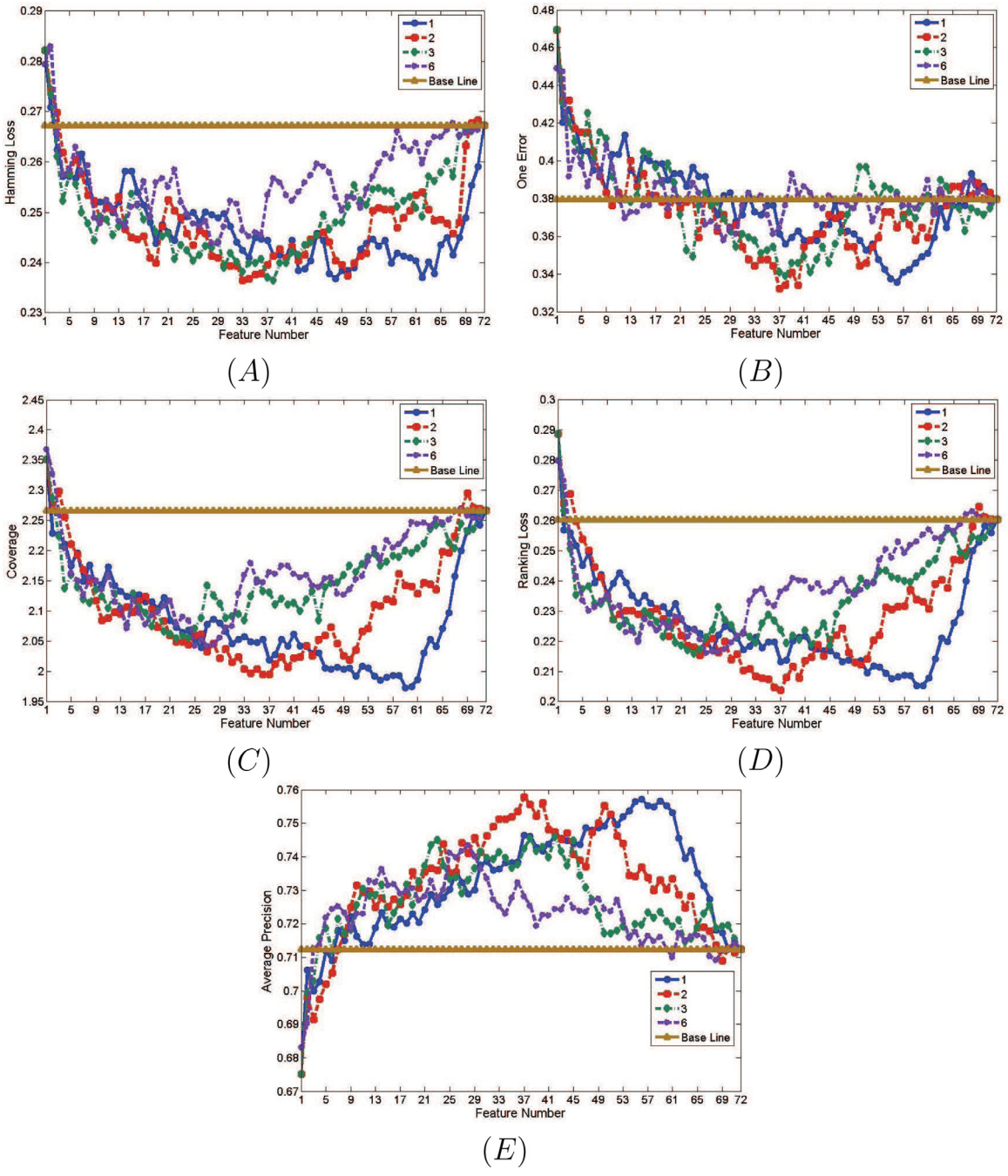

Fig. 5. Classification performance on *Emotions* according to the number of selected features using the proposed method with different numbers (1,2,3,6) of information granules: (*A*) Hamming loss; (*B*) One error; (*C*) Coverage; (*D*) Ranking loss; (*E*) Average precision.

ground-truth labels of the instance.

$$\gamma_{Covergage} = \frac{1}{t}\sum_{i=1}^{t}\max_{l\in Y_i} r_i(l) - 1, \tag{18}$$

where $r_i(l)$ is a ranking function, which maps the predicted outputs for any $l \in Y_i$. The labels of each test instance are ranked in descending order according to the associated probability. The smaller the value of the coverage, the better the performance.

(3) *One-error*: it computes how many times the top-ranked label is not in the set of relevant labels of the instance. The smaller the value of one error, the better the performance.

$$\gamma_{Oneerror} = \frac{1}{t}\sum_{i=1}^{t}\langle[\arg\max_{l\in L} r_i(l)] \notin Y_i\rangle. \tag{19}$$

(4) *Ranking Loss*: it expresses the number of times that irrelevant labels are ranked higher than relevant labels.

$$\gamma_{RankingLoss} = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{|Y_i||\overline{Y}_i|}|\{(l_a, l_b) : r_i(l_a) > r_i(l_b), (l_a, l_b) \in Y_i \times \overline{Y}_i\}|, \tag{20}$$

where $\overline{Y}_i$ denotes the complementary set of Y_i set which is the complement with respect to L. The smaller the value of one error, the better the performance.

(5) *Average Precision*: the criterion evaluates the average fraction of labels ranked above a particular label $l \in Y_i$, which actually are in Y_i. The bigger the value of average precision, the better the

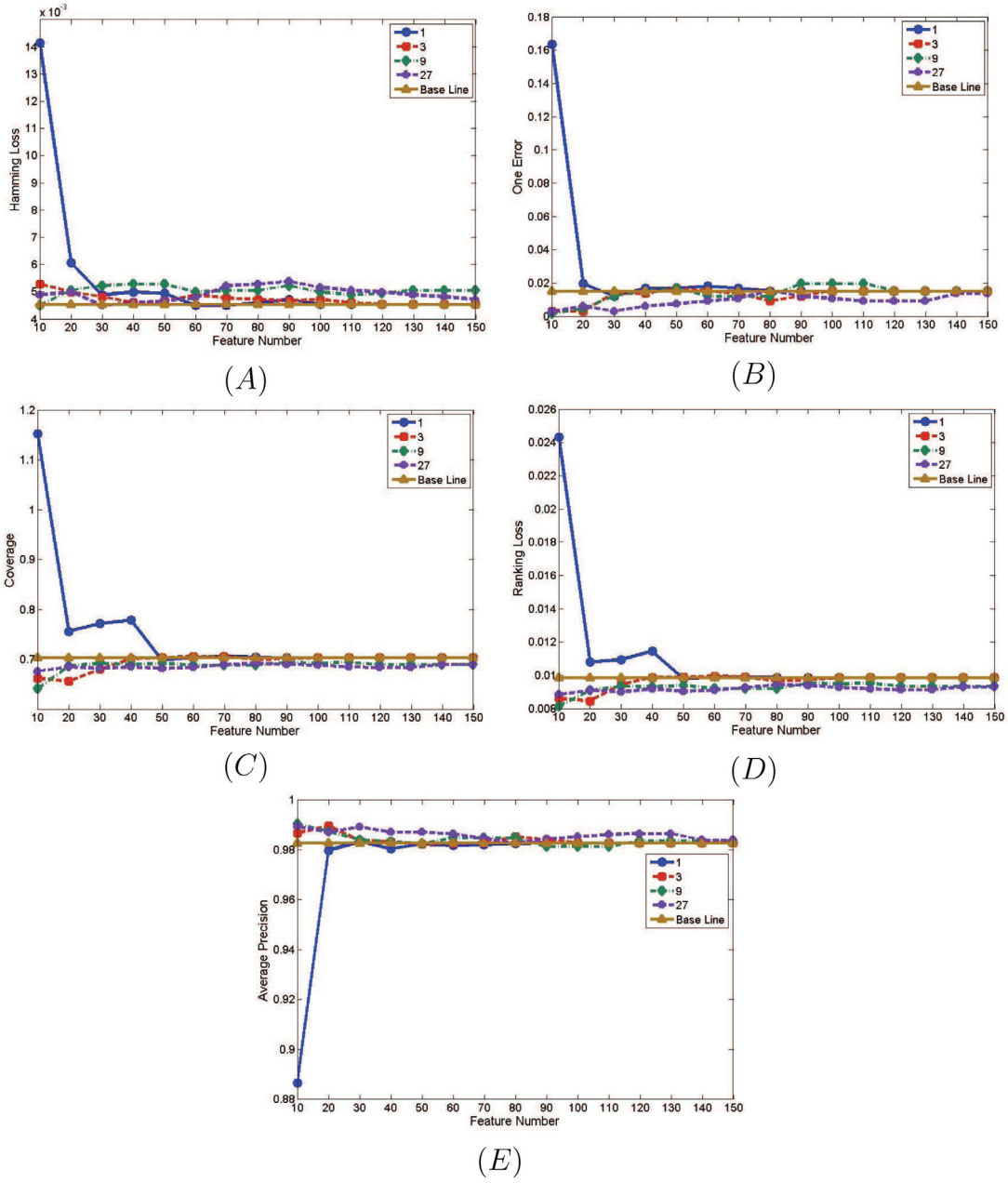

Fig. 6. Classification performance on *Genbase* according to the number of selected features using the proposed method with different numbers (1,3,9,27) of information granules: (*A*) Hamming loss; (*B*) One error; (*C*) Coverage; (*D*) Ranking loss; (*E*) Average precision.

performance.

$$\gamma_{AveragePrecision} = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{|Y_i|}\sum_{l\in Y_i}\frac{|\{l' \in Y_i : r_i(l') \le r_i(l)\}|}{r_i(l)}. \quad (21)$$

5.3. Granulation test

In this section, the effect of the label granularity is researched. As shown in Figs. 3, 4, 5 and 6, we involve a base line in each figure denoting the direct results of the use of the ML-KNN without any preprocessing of features or labels. Furthermore, in Figs. 3, 4, 5 and 6, the method with one information granule means no granulation processing, while the method with the largest number of information granules converts the label set into numerous single labels and shows the performance without considering the label dependency. More specifically, 1, 9, 27, and 53 label information granules are considered for *Enron*, and 1, 5, 9, and 45 label information granules are studies in case of *Medical*, while 1, 2, 3, and 6 label information granules for *Emotions*, and 1, 3, 9, and 27 label information granules for *Genbase*. Different numbers of features are chosen, which can demonstrate the relationship between the feature number and the granularity of the label information granule. Figs. 3, 4, 5 and 6 show the performances obtained for the data set *Enron, Medical, Emotions*, and *Genbase*, respectively.

In Fig. 3, the feature numbers are selected from 10 to 150 with a step of 10, and 200, 250, 300. It can be seen that almost all methods with feature selection achieve better results than the base method on all set feature number, except the one with one information granule, because the feature number is too small resulting in a significant loss of information. However, the performance of the method with one information granule improves quickly with the increasing of feature number, and performs better than the

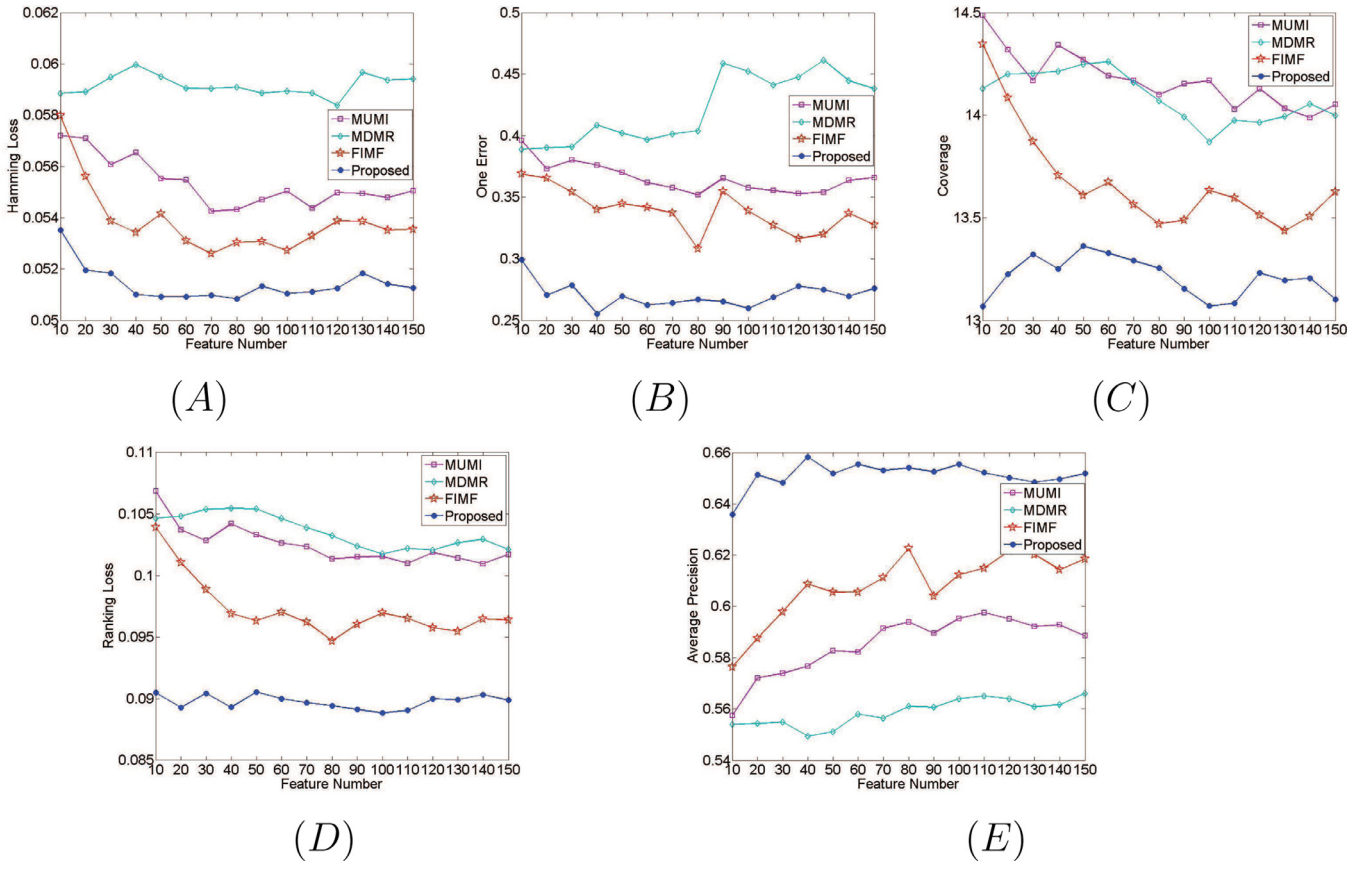

Fig. 7. Classification performance on *Enron* according to the number of selected features using compared feature selection methods: (*A*) Hamming loss; (*B*) One error; (*C*) Coverage; (*D*) Ranking loss; (*E*) Average precision.

base method when the feature number is greater than 120. The performance of the method with a single information granule is enhanced as the feature number increases. However, the performances of the four methods endowed with information granulation increase reaching the best performance, and then decrease gradually with the increasing of the number of features. The methods with granulation perform better than the one without granulation. Furthermore, the method with smaller size of information granule earlier reaches the best performance. As shown in Fig. 3, for *Enron*, the method with 27 information granules results in the best performance with regard to the Hamming loss criterion and one error criterion, and the performance is very close to the best one, obtained by using the method with 53 information granules, on the other three criteria. The label dependency is an important factor, although the method considering the label dependency does no better than the method transforming the label set into single labels with regard to some criteria because of the loose label dependency observed for the *Enron* data set. Thus, the feature selection and information granulation can improve the performance of the method for this data set.

With regard to the *Medical* data set, the size of selected feature subset ranges from 10 to 150 with a step of 10, with the performance displayed in Fig. 4. Compared to the base method, the performance shape of *Medical* is similar to the one of *Enron* on Hamming loss, one error and average precision. Regarding coverage and ranking loss, the base method is worse than the best performance of the methods taking the label dependency into consideration, namely the method with 1, 5, and 9 information granules. From Fig. 4, it can be seen that the feature number with the best performance has a negative correlation with the number of information granules, and after obtaining the best performance, the performance of each considered method degrades gradually as the number of features increases. For *Medical*, the best performance is achieved by the method with 9 information granules on one error and average precision, while the method with 5 information granules outperforms on the rest of the criteria. Furthermore, the best performances of the method with 5, 9 information granules are greater than the performances of the method with 1, 45 information granules. Therefore, the proposed method performs well on *Medical* data set.

For *Emotions* data set, all feature numbers are considered, because it has a small number of features as shown in Table 1. From Fig. 5, it can be seen that all feature selection methods firstly vary forward to the best performance, then change away as the feature number increases, and finally achieve the same results as the base method when all of the 72 features are selected. All of the methods considering feature selection always perform significantly better than the base method except the beginning and the ending, because very few features are selected, and the feature subset does not have enough class-discriminative information, or too many features are selected, and irrelevant features degrade the performance. The best performances on *Emotions* are obtained by the method with 2 information granules on all evaluation criteria beside the coverage. The method with 2 information granules gets the best performance around 37 features and is earlier than the one with 1 information granule, which reaches the best around 60 features, namely the best feature subset of the former is smaller than the later, because the label information granules with a coarser granularity has a larger best feature subset. The method with 6 information granules perform not very well, because of ignoring label dependency.

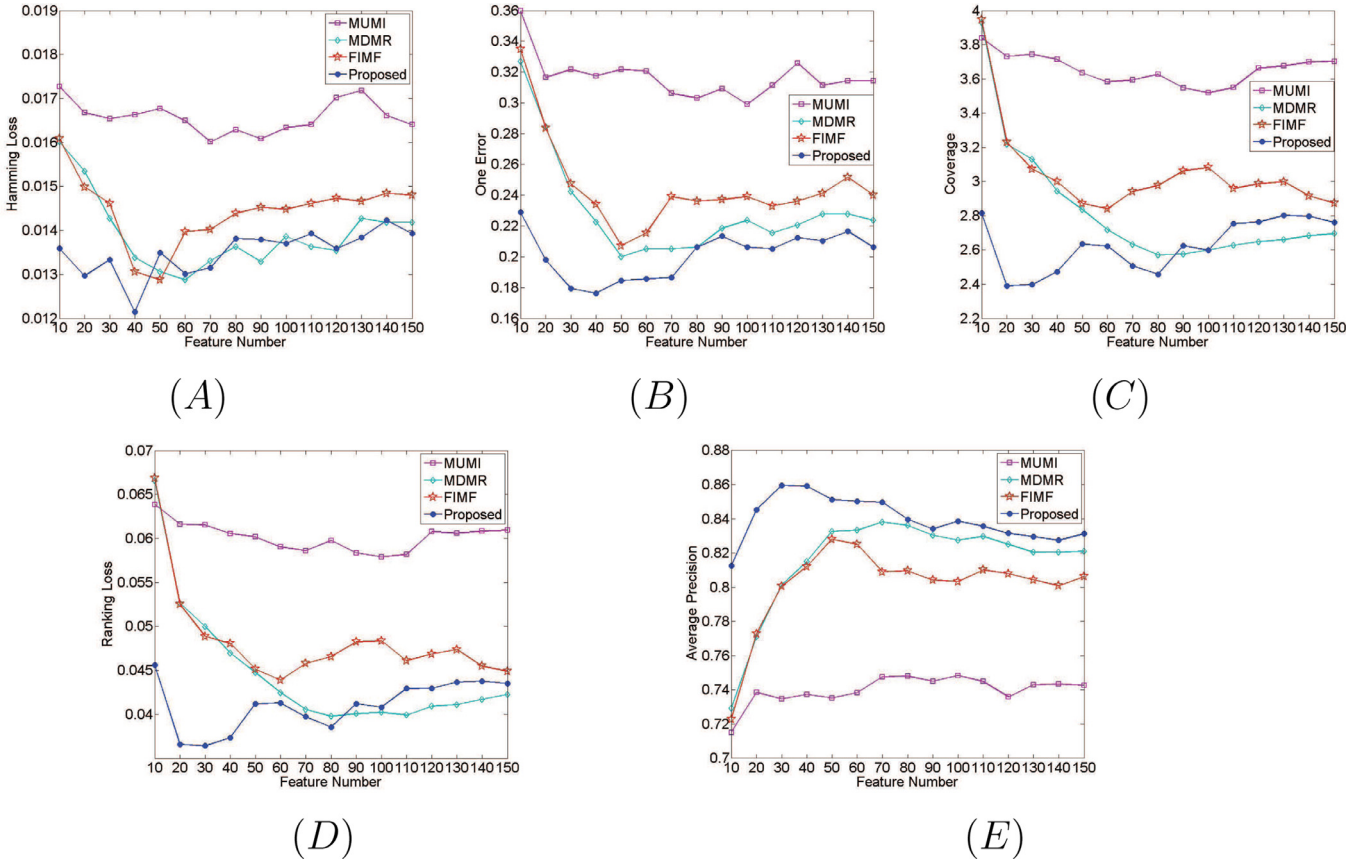

Fig. 8. Classification performance on *Medical* according to the number of selected features using compared feature selection methods: (*A*) Hamming loss; (*B*) One error; (*C*) Coverage; (*D*) Ranking loss; (*E*) Average precision.

As shown in Fig. 6, all the performances on *Genbase* are very excellent versus the feature number which is set from 10 to 150 with a step of 10, for the *Genbase* data set has a few labels and very small label cardinality, while many features. Compared with the base method, the methods considering feature selection still perform better, although the base method obtains very excellent performance. The method with 9 information granules outperforms among all the methods. The method without granulating achieves stable results after 50 features.

From the experiment, the proposed feature selection method can improve the algorithm performance of ML-KNN, and performs better than those without granulation or ignoring label dependency. There are two properties summarized about the granular feature selection in multi-label learning as follows:

- The size of selected feature subset influences the algorithm performance. For a given label information granule, the selected feature subset is too small resulting in losing much information, not able to stand for the class-discriminative information contained in the data set. If too great features are selected, the unnecessary feature do not only increase the computing cost, but also may degrade classification accuracy.
- The label granularity affects the size of selected feature subset achieving the best performance. The label granularity is more rough, namely more labels are considered in one label information granule, less information losing about label dependency. However, the selected feature subset is large for the label information granule with rough granularity, and the selected feature probably has significant dependency on some labels in the label information granule, while little, even negative effect on others, so that the algorithm performance is degraded.

Furthermore, there is a drawback on the method without considering the label dependency that the algorithm complexity is much higher than others because of training a classification model for each label information granule.

5.4. Comparative analysis

To show the quality of the proposed method compared to others, we compare the proposed method with multi-label feature selection methods MUMI [21], MDMR [22] and FIMF [37] which are introduced in Section 2. The compared results on *Genbase* are not displayed in this section, because it is very easy to classify, and the results cannot show the differences of the compared methods. Considering the results in Section 5.3, for multi-label data sets *Enron, Medical* and *Emotions*, the number of information granules is correspondingly set to be 9, 5 and 2 on the proposed method. The compared results are shown in Figs. 7, 8 and 9.

From Figs. 7, 8 and 9, it is intuitively noticed that for a single data set, the compared performances in each evaluation criteria has similar variation, so we don't discuss the individual result in every evaluation criterion. As shown in Fig. 7, the proposed method outperforms the others on each considered feature number on *Enron*. On *Medical*, it can be seen from Fig. 8, that the proposed method obtains the best performance on a smaller feature number than others, and the obtained best performance is better than other three. It means that the proposed method is fast and effective. For the experiment on *Emotions* displayed in Fig. 9, the proposed method achieves much better results than MUMI. Compared to MDMR and FIMF, the proposed method has a similar variation as the feature number increases, and even gets a little better results on some evaluation criteria. Among the three data sets, the

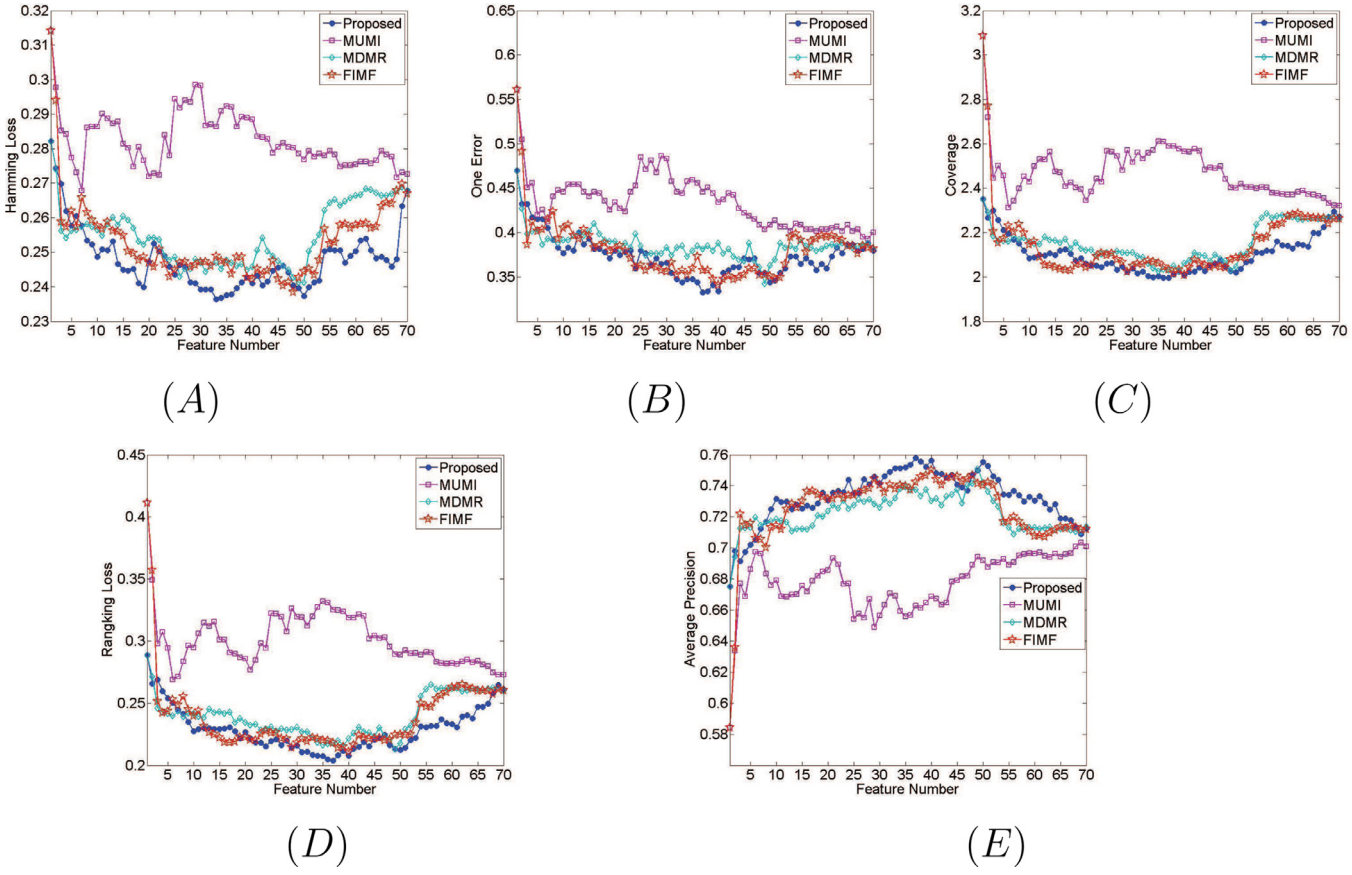

Fig. 9. Classification performance on *Emotions* according to the number of selected features using using compared feature selection methods: (*A*) Hamming loss; (*B*) One error; (*C*) Coverage; (*D*) Ranking loss; (*E*) Average precision.

Emotions data set has the smallest label number, and it cannot exert the advantage of the proposed method very well. Overall, the proposed method can get a better quality than MUMI, MDMR and FIMF on the used data sets, particularly on data sets with a larger number of labels.

6. Conclusions

In this paper, we have proposed a granular feature selection method for multi-label learning with a maximal correlation and minimal redundancy criterion, which takes the local label dependency into consideration. The maximal correlation and minimal redundancy criterion can approximate the maximal dependency on feature selection in multi-label learning, where the selected features are most relevant to the target label information granule, meanwhile there is the least redundancy in the feature subset. To verify the effectiveness, the performance of the proposed method has been compared with those not reducing the features, not granulating the labels, or not exploring the label dependency. We also have compared the proposed method and three well known multi-label feature selection methods. The experimental results show that the proposed method performs better and can select a compact feature subset. Especially, the multi-label feature selection could have very significant impact on clinical application, including one emerging field named radiomics, where a large number of quantitative imaging features have been fitted into a computational model to predict the clinical interested outcomes. Robust and meaningful feature selection algorithms can largely improve the model performance in this field. Therefore, it is necessary to explore the label dependency properly with granular computing in multi-label feature selection.

Although the proposed method can reduce the dimensionality of the feature space, it requires the training of the classifier for every label information granulation. This may substantially increase the algorithm complexity, when the number of information granule is very large. Hence, it is worthwhile to research and develop a way of measuring a granularity to guide on how to determine the best number of information granules.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Serial No. 61673301, 61273304, 61573259, 61573255), and the Specialized Research Fund for the Doctoral Program of Higher Education of China (File No. 20130072130004).

Appendix

Section 4.2 shows that the maximal correlation of feature x_i and information granule G can be approximated by maximizing the dependency of x_i and each label l_j in G. Firstly, considering the maximal correlation criterion in Eq. (12), the mutual information between feature x_i and information granule G can be represented in terms of entropies through Lemma 1 as follows:

$$MI(x_i; G) = H(x_i) + H(G) - H(x_i, G). \tag{22}$$

Science the labels in information granule G are fixed, it can be easily seen that the entropy of G is a constant. Using Eqs. (12) and

(22), we have

$$\begin{aligned}\max MI(x_i; G) &= \max(H(x_i) + H(G) - H(x_i, G)) \\ &= \max\ (H(x_i) - H(x_i, G)) + H(G) \\ &\propto \max(H(x_i) - H(x_i, G)).\end{aligned} \tag{23}$$

Here, we define a quantity $\Theta(x_i; G)$ representing the term $H(x_i)$ minus $H(x_i, G)$. Then, the upper bound of $\Theta(x_i; G)$ can be formed as follows:

$$\begin{aligned}\Theta(x_i; G) &= H(x_i) - H(x_i, G) \\ &= -\sum p(x_i) log p(x_i) + \sum \cdots \sum p(x_i, l_1, l_2, \cdots, l_r) \\ &\quad \times log p(x_i, l_1, l_2, \cdots, l_r) \\ &= \sum \cdots \sum p(x_i, l_1, l_2, \cdots, l_r) log \frac{p(x_i, l_1, l_2, \cdots, l_r)}{p(x_i)} \\ &= \sum \cdots \sum p(x_i, l_1, l_2, \cdots, l_r) \\ &\quad \times log \frac{p(l_1|l_2, \cdots, l_r, x_i)\, p(l_2|l_3, \cdots, l_r, x_i) \cdots p(l_r|x_i)\, p(x_i)}{p(x_i)} \\ &= -H(l_1|l_2, \cdots, l_r, x_i) - H(l_2|l_3, \ldots, l_r, x_i) - \cdots - H(l_r|x_i) \\ &\leq 0.\end{aligned} \tag{24}$$

Obviously, the upper bound of $\Theta(x_i; G)$ is 0 while it obtains the maximal value when all variables are maximally dependent. In Eq. (24), for labels in G are already given and this means that x should have the maximal correlation on each label l_j in G, i.e. the condition shown in Eq. (13).

References

[1] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, in: Data Mining and Knowledge Discovery Handbook, Springer, 2010, pp. 667–685.

[2] G. Tsoumakas, I. Katakis, Multi label classification: an overview, Int. J. Data Warehouse. Min. 3 (3) (2007) 1–13.

[3] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.

[4] G. Madjarov, D. Kocev, D. Gjorgjevikj, S. Dzeroski, An extensive experimental comparison of methods for multi-label learning, Pattern Recognit. 45 (9) (2012) 3084–3104.

[5] Y. Yu, W. Pedrycz, D.Q. Miao, Neighborhood rough sets based multi-label classification for automatic image annotation, Int. J. Approx. Reason. 54 (2013) 1373–1387.

[6] R.E. Schapire, Y. Singer, Boostexter: a boosting-based system for text categorization, Mach. Learn. 39 (2000) 135–168.

[7] P. Pavlidis, J. Weston, J. Cai, W.N. Grundy, Combining Microarray Expression Data and Phylogenetic Profiles to Learn Functional Categories Using Support Vector Machines, in: Proceedings of the Fifth Annual International Conference on Computational Biology, Montreal, Canada, 2001, pp. 242–248.

[8] J. Wu, T. Aguilera, D. Shultz, et al., Early-stage non-small cell lung cancer: quantitative imaging characteristics of (18)f fluorodeoxyglucose PET/CT allow prediction of distant metastasis, Radiology 281 (1) (2016a) 270–278.

[9] J. Wu, G. Gong, Y. Cui, et al., Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy, J. Magnet. Resonance Imag. Jmri 44 (5) (2016b) 1107–1115.

[10] J. Wu, M.F. Gensheimer, X. Dong, et al., Robust intratumor partitioning to identify high-risk subregions in lung cancer: a pilot study, Int. J. Radiat. Oncol. Biol. Phys. 95 (5) (2016c) 1504–1512.

[11] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[12] M.L. Bermingham, et al., Application of high-dimensional feature selection: evaluation for genomic prediction in man, Sci. Rep. 5 (2015).

[13] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[14] W.Z. Chen, J. Yan, B.Y. Zhang, Q. Yang, Document transformation for multi-label feature selection in text categorization, in: Proceeding of the Seventh IEEE International Conference on Data Mining, 2007, pp. 451–456.

[15] K. Trohidis, G. Tsoumakas, G. Kalliris, I. Vlahavas, Multilabel classification of music into emotions, in: Proceeding of the Ninth International Conference of Music Information Retrieval, 2008, pp. 325–330.

[16] M. Valipour, Optimization of neural networks for precipitation analysis in a humid region to detect drought and wet year alarms, Meteorol. Appl. 23 (1) (2016) 91–100.

[17] M. Valipour, M.A.G. Sefidkouhi, S. Eslamian, Surface irrigation simulation models: a review, Int. J. Hydrol. Sci. Technol. 5 (1) (2015) 51–70.

[18] S.I. Yannopoulos, G. Lyberatos, N. Theodossiou, et al., Evolution of water lifting devices (pumps) over the centuries worldwide, Water (Basel) 7 (9) (2015) 5031–5060.

[19] M. Zhang, J.P. Peña, V. Robles, Feature selection for multi-label naive bayes classification, Inf. Sci. 179 (2009) 3218–3229.

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

[21] J. Lee, D.W. Kim, Feature selection for multi-label classification using multivariate mutual information, Pattern Recognit. Lett. 34 (2013) 349–357.

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

[23] J. Pestian, C. Brew, P. Matykiewicz, et al., A shared task involving multi-label classification of clinical free text, in: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing (BioNLP'07), Association for Computational Linguistics, Stroudsburg, PA, USA, 2007, pp. 97–104.

[24] W. Pedrycz, Granular computing: an introduction, in: IFSA World Congress and 20th NAFIPS International Conference, 2001, pp. 1349–1354. Joint 9th

[25] W. Pedrycz, B.J. Park, S.K. Oh, The design of granular classifiers: A study in the synergy of interval calculus and fuzzy sets in pattern recognition, Pattern Recognit. 41 (12) (2008) 3720–3735.

[26] Y.Y. Yao, Interpreting concept learning in cognitive informatics and granular computing, IEEE Trans. Syst. Man Cybern. Part B 39 (4) (2009) 855–866.

[27] Y.Y. Yao, L.Q. Zhao, A measurement theory view on the granularity of partitions, Inf. Sci. (Ny) 213 (2012) 1–13.

[28] C. Park, M. Lee, On applying linear discriminant analysis for multi-labeled problem, Pattern Recognit. Lett. 29 (7) (2008) 878–887.

[29] M.R. Boutell, J. Luo, X. Shen, et al., Learning multi-label scene classification, Pattern Recognit. 37 (9) (2004) 1757–1771.

[30] G. Tsoumakas, I. Vlahavas, Random k-labelsets: An ensemble method for multilabel classification, in: Machine learning: ECML'07, Springer, Berlin, 2007, pp. 406–417.

[31] J. Read, A pruned problem transformation method for multi-label classification, in: Proceeding of New Zealand Computer Science Research Student Conference, 2008, pp. 143–150.

[32] M. Valipour, Sprinkle and trickle irrigation system design using tapered pipes for pressure loss adjusting, J. Agric. Sci. 4 (12) (2012a) 125–133.

[33] M. Valipour, Comparison of surface irrigation simulation models: full hydrodynamic, zero inertia, kinematic wave, J. Agric. Sci. 4 (12) (2012b) 68–74.

[34] M.M. Khasraghi, M.A.G. Sefidkouhi, M. Valipour, Simulation of open- and closed-end border irrigation systems using SIRMOD, Arch. Agron. Soil Sci. 61 (7) (2015) 929–941.

[35] N. Spolaôr, E.A. Cherman, M.C. Monard, H.D. Lee, A comprasion of multi-label feature selection methods using the problem tranformation approch, Electron. Notes Theor. Comput. Sci. 292 (2013) 135–151.

[36] G. Tsoumakas, I. Katakis, I. Vlahavas, Random k-labelsets for multilabel classification, IEEE Trans. Knowl. Data Eng. 23 (7) (2011) 1079–1089.

[37] J. Zhang, D.W. Kim, Fast multi-label feature selection based on information-theoretic feature ranking, Pattern Recognit. 48 (2015) 2761–2771.

[38] C. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948a) 379–423.

[39] C. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (4) (1948b) 623–656.

[40] A.D. Wyner, A definition of conditional mutual information for arbitrary ensembles, Inf. Control 38 (1) (1978) 51–59.

[41] G. Tsoumakas, I. Katakis, I. Vlahavas, Effective and efficient multilabel classification in domains with large number of labels, in: Proceeding of ECML/PKDD 2008 Workshop on Mining Multidimensional Data, 2008, pp. 287–313.

[42] G. Tsoumakas, E. Spyromitros-Xiousfis, I.V.J. Vilcek, Mulan:a java library for multi-label learning, J. Mach. Learn. Res. 12 (2011) 2411–2414.

[43] http://bailando.sims.berkeley.edu/enron_email.html.

[44] S. Diplaris, G. Tsoumakas, P. Mitkas, et al., Protein classification with multiple algorithms, in: Proceedings of the 10th Panhellenic Conference Informatics, PCI 2005, Volas, Greece, 2005, pp. 448–456.

[45] M.L. Zhang, Z.H. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048.

[46] R.E. Schapire, Y. Singer, Boostexter: a boosting-based system for text categorization, Mach. Learn. 39 (2) (2000) 135–168.

Feng Li is a Ph.D candidate with the department of computer science and technology of Tongji University, Shanghai, China. His researches focus on multi-label learning and granular computing.

Duoqian Miao received the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1997. He is currently a Professor with the School of Electronics and Information Engineering and the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China. His current research interests include soft computing, rough sets, pattern recognition, data mining, machine learning, and intelligent systems.

Witold Pedrycz received the M.Sc., Ph.D., and D.Sci. degrees from Silesian University of Technology, Gliwice, Poland, in 1977, 1980, and 1984, respectively. He is currently a Professor and Canada Research Chair (CRCcomputational intelligence) with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. His current research interests include computational intelligence, fuzzy modeling and granular computing, knowledge discovery and data mining, fuzzy control, pattern recognition, knowledge-based neural networks, relational computing, and software engineering. He has published numerous papers in this area.