Knowledge-Based Systems 195 (2020) 105684

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Mutual information-based label distribution feature selection for multi-label learning☆

Wenbin Qian [a,b], Jintao Huang [c], Yinglong Wang [c,*], Wenhao Shu [d]

[a] School of Software, Jiangxi Agricultural University, Nanchang 330045, China
[b] Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China
[c] School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang 330045, China
[d] School of Information Engineering, East China Jiaotong University, Nanchang 330013, China

ARTICLE INFO

Article history:
Received 30 September 2019
Received in revised form 18 February 2020
Accepted 20 February 2020
Available online 25 February 2020

Keywords:
Feature selection
Multi-label data
Granular computing
Label enhancement
Mutual information

ABSTRACT

Feature selection used for dimensionality reduction of the feature space plays an important role in multi-label learning where high-dimensional data are involved. Although most existing multi-label feature selection approaches can deal with the problem of label ambiguity which mainly focuses on the assumption of uniform distribution with logical labels, it cannot be applied to many practical applications where the significance of related label for every instance tends to be different. To deal with this issue, in this study, label distribution learning covered with a certain real number of labels is introduced to design a model for the labeling-significance. Nevertheless, multi-label feature selection is limited to handling only labels consisting of logical relations. In order to solve this problem, combining the random variable distribution with granular computing, we first propose a label enhancement algorithm to transform logical labels in multi-label data into label distribution with more supervised information, which can mine the hidden label significance from every instance. On this basis, to remove some redundant or irrelevant features in multi-label data, a label distribution feature selection algorithm using mutual information and label enhancement is developed. Finally, the experimental results show that the performance of the proposed method is superior to the other state-of-the-art approaches when dealing with multi-label data.

1. Introduction

In traditional machine learning, single-label learning [1–4] handles the problem that one sample remains associated only with a relevant candidate label, which is denoted as single-label classification [5–7]. However, in practical situations, there are many tasks where one sample has the possibility to be associated with multiple labels simultaneously, which is called multi-label learning [8–10]. For instance, a news may be viewed as the social, political and martial. A piece of music may be related to sadness and madrigal. An image may be described by mountains, water bodies or trees. Multi-label learning is a research hot spot for practical applications. For example, a class of generative statistical topic models is investigated in [11]. A representation model based on boosting to denote text documents is proposed in [12]. A new approach for image annotation is proposed in [13]. A novel method for multi-label classification using the multi-label twin support vector machine is proposed in [14]. A new approach namely multi-label learning with emerging new labels is proposed in [15] to detect and classify those instances with emerging new labels. A novel sparse and low-rank representation-based method is proposed to exploit the asymmetric correlations for multi-label classification in [16].

In multi-label learning, the significance of each label is regarded as equivalent, i.e., logical labels with two logical relationships between yes and no. Whether the target label y can describe the sample x correctly, a general method is performed where one sample x can be assigned with a possible label y for $L_x^y \in \{0, 1\}$. Single-label learning and multi-label learning only answered question of which label should the sample belong to, but not the relative significance of different labels for each sample. However, in various practical applications, neither single-label learning nor multi-label learning can deal with the problem of different significance for each label, i.e., various labels may be related to one sample with different significance. Taking a facial expression such as emotion analysis as an example, a facial expression often conveys a complex mixture of multiple emotions

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2020.105684.

* Corresponding author.
E-mail address: wangylx@126.com (Y. Wang).

https://doi.org/10.1016/j.knosys.2020.105684
0950-7051/

(e.g., happiness, sadness, surprise) [17]. The various intensities of all the basic emotions naturally form an emotion distribution for the facial expression. The emotion expressions with the highest intensity or emotions with higher intensities than a threshold as the positive labels can be fit into the single-label learning or multi-label learning framework. Unfortunately, the significant information of different intensities of the related emotions will be lost. In reality, it is increasingly common for such data to have a different relative significance for each label. Therefore, it is reasonable to use a more adaptive label description rather than a hard labeled one. In order to solve this problem, Geng proposed a generalized machine learning paradigm named label distribution learning [17–19], which can deal with the problem "how much does each label describe the sample". For one sample x, a real number d_x^y could be assigned to each possible label $y_j \in \{y_1, y_2, \ldots, y_m\}$ and it is regarded as the degree of description $d_x^{y_j} \in [0, 1]$, $\sum_{j=1}^{m} d_x^{y_j} = 1$. Despite various practical applications in label distribution learning, the high-dimensional feature space existing in label distribution learning has not been exploded.

As it is known that for a given learning task, there are many redundant or irrelevant features, which will bring several disadvantages to learning algorithm. Therefore, before using a data set, it is necessary to preprocess the data to remove redundant features. Feature selection, also known as attribute reduction, has been regarded as a significant step towards dimensionality reduction and improving classification performance [7,20]. It is known that feature selection is an effective measurement to reduce the high-dimensionality in multi-label learning. Differed from single-label feature selection, multi-label feature selection should consider the relevance of multiple labels rather a single label. The existing multi-label feature selection methods are generally classified into three groups, including filter [21,22], wrappers [23, 24] and embedded methods [25,26]. In the filter methods, feature subsets are independent of any machine learning algorithm, and make use of some criteria or scores to eliminate the features. In addition, it becomes the most suitable method for feature selection because of its low computational complexity. The wrapper method considers the accuracy of a classification algorithm for feature selection. Although a better classification performance can be achieved, the computation is high. The embedded method exploits the advantages of both the filter and wrappers methods. Here, the process of obtaining an efficient feature is embedded in the classification process.

Several researches on multi-label feature selection have been exploited. Nevertheless, the significance of related labels for each instance is not considered for multi-label learning. In other words, feature selection for multi-label data with logical labels is not efficiently handled. In many real-world applications, the significance of related labels for every instance is often different. Most of existing multi-label feature selection algorithms focus on the same label significance, which may affect the classification results. To address with issue, label distribution learning is used for multi-label feature selection in this paper. The two main contributions can be summarized as follows: Firstly, a label-enhancement algorithm using random variable distribution and granular computing is proposed. This algorithm converts the traditional logical labels in multi-label data into the labels with different significance namely label distribution, which reflects the significance of the corresponding labels. Second, to remove some redundant features effectively, a label distribution feature selection algorithm based on mutual information is presented, which solves the problem of high dimensionality in label distribution learning.

The rest of this paper is structured as follows: some related work on label distribution learning and multi-label feature selection are reviewed in Section 2. A label enhancement algorithm based on a label distribution decision system is proposed in Section 3. Based on the mutual information, a label distribution feature selection algorithm is developed in Section 4. To validate the proposed methods, the experimental results on different real data sets are presented in Section 5. The conclusions and outline plans for further research are drawn in Section 6.

2. Related work

2.1. Multi-label feature selection

Multi-abel feature selection has been applied successfully in recent years, many efforts have been made to select feature subset for multi-label data. In multi-label learning, feature selection is a significant step to mitigate the curse of dimensionality in high-dimensional data [27–29]. Several feature selection algorithms have been designed to maintain or improve the accuracy for multi-label learning with three categorized types, i.e., semi-supervised, supervised and unsupervised measures. In terms of supervised measures in multi-label learning, Lin et al. introduced the fuzzy mutual information to evaluate the quality of features in multi-label learning and designed an efficient algorithm to conduct multi-label feature selection when the feature space is known in advance [30], completely or partially. Lin et al. proposed a novel fuzzy rough set model for multi-label feature selection [21] to search the real different classes' samples for the target sample, with effectiveness of the robustness of upper and lower fuzzy approximations. Li et al. proposed a maximal correlation minimal redundancy criterion based on the mutual information that ensures the selected feature subset contains the most class-discrimination information [31]. Yu et al. explored a multi-label linear discriminant analysis for multi-label dimension reduction [32]. Lim et al. designed a score function using mutual information and proposed a numerical optimization approach to avoid being stuck in the local optima [33]. Xu et al. proposed a FRS-LIFT method to solve the problem of the increasing feature dimensionality and a series of redundant information existing in feature space, which implements label-specific feature selection with a fuzzy system [34]. Wang et al. proposed a new hybrid feature selection using multi-filter weights and multi-feature weights [35]. For semi-supervised measures, Mikalsen et al. proposed a novel semi-supervised and multi-label dimensionality reduction method that effectively utilizes information from both noisy multi-labels and unlabeled data [29]. Juan et al. presented an empirical evaluation of feature selection techniques in the context of three main MLC transformation methods [36]. Wang et al. proposed a novel semi-supervised multi-label feature selection algorithm and apply it to three different applications: natural scene classification, web page annotation, and gene functional classification [37]. Wang et al. defined the feature interaction to select more valuable features for missing labels, which may be ignored because of the incomplete label space [38].

And with respect to unsupervised measures, some research has performed to solve the issue of missing labels, respectively. For instance, Zhu et al. proposed a novel model for multi-label feature selection under the circumstance of missing labels [26]. Liu et al. has used SVM to process the missing labels only with the partial available labels for multi-label learning [13]. Ma et al. recovered the label by performing the label matrix imputation in the labeled space simultaneously [39]. He et al. proposed a unified learning framework for multi-label classification joint with label correlations, missing labels, and feature selection to solve the challenges existing in multi-label learning [24]. Chen et al. solved the optimization problem by proposing a multi-class and multi-label feature selection algorithm using an adaptive least absolute shrinkage and selection operator [40].

Additionally, the online streaming multi-label feature selection has attracted by many scholars. For instance, Liu et al. proposed an online multi-label group feature selection with two phases namely online group selection and the online inter-group selection to deal with the problem of considering group structures of features intrinsically [41]. Liu et al. presented the online attribution reduction framework for streaming multiple label data based on the maximum-nearest-neighbor granularity [42]. Kashef et al. [43] designed a fast and accurate multi-label feature selection algorithm based on the filter method to find label-specific features, which can be used for online feature selection simultaneously.

For the aforementioned applications, the multi-label data problems such as the label correlations [8,23,44], missing labels [13,24,26], high-dimensionality [21,45], streaming features [30], and label-specific features [23,34,46] have been addressed by scholars. Nevertheless, multi-label learning can only answer the question "which label can describe the instance" rather than handle the question with more ambiguity "how much does each label describe the instance" directly, i.e., the mentioned researches have not involved the relative significance of each label in the description of each instance [17]. In reality, the relative significance of each label is more common for the real-world data. To address this issue, label distribution learning is introduced to promote the learning task more effective in multi-label learning [17,47].

2.2. Label distribution learning

The emergence of label distribution learning makes it possible to learn more semantics from data than multi-labels. Some real applications based on the label distribution have been investigated. Chen et al. proposed a method called structured random forest (StructRF) for semantic image labeling and edge detection, which treats the distribution as an integral whole to make full use of the structural information among different classes [48]. Zhang et al. introduced the label distribution learning framework into crowd counting with the labeled crowd images for label distributions instead of the conventional single labels [49]. It is known that the label correlation is important for label distribution learning. To learn and leverage the label correlations simultaneously, Zhao et al. proposed a new approach based on the optimal transport theory by learning the transportation and ground metric [50]. Geng et al. regarded the face image as a sample related to the label distribution for facial age estimation, which not only learn the chronological age, but also its adjacent ages [47,51]. Xu et al. solved the problem of incomplete supervised information using label distribution and proposed a proximal gradient descend algorithm to exploit the correlation between labels based on trace norm minimization [52]. Gao et al. converted the image labels into a label distribution, which utilized the feature learning and classifier learning for label ambiguity [53]. Yang et al. introduced a novel knowledge distilling strategy to assist visual feature learning in the convolution neural network for painting style classification [54]. Liu et al. introduced a strategy for learning label distributions with only five-to-ten labels per item [55] by aggregating human-annotated labels over multiple.

For all practical purposes, single-label learning and multi-label learning can be regarded as the specific cases of label distribution learning. Consequently, there are some important theories and applications for investigating this machine learning paradigm. Nevertheless, existing researches on label distribution learning almost concentrated on the classification task; however, the dimensionality of feature space have not been investigated. It tends to be over-fit when learning a model with several features, which can result in reduced performance on unseen data. The high-dimensional data can significantly increase the computational costs for data analysis and the memory storage requirements [6,56,57]. Therefore, feature selection, a component of dimensionality reduction, is applied to the label significance generated by label distribution learning to select a more efficient feature subset for multi-label classification, which can improve the learning performance, build a better model and increase the computational efficiency.

3. Label enhancement for multi-label neighborhood decision system

In this section, we introduce the multi-label neighborhood decision system, and subsequently propose a label enhancement algorithm based on the random variable distribution.

3.1. Multi-label neighborhood decision system

Some basic concepts in rough set-based granular computing are reviewed in this subsection [58]. We assume the universe U is a finite nonempty set, $U/IND(B)$ is a partition of U induced by the feature subset B, for $\forall x \subset U$, $[x]_B$ is the class containing x in $U/IND(B)$ [59,60].

Definition 1. Let $U = \{x_1, x_2, \ldots, x_n\}$ be a nonempty finite set of objects, $C = \{c_1, c_2, \ldots, c_p\}$ be a nonempty finite set of conditional features with p features and $L = \{l_1, l_2, \ldots, l_m\}$ be a nonempty finite set of labels with m possible labels. A multi-label decision system (MLDS) can be represented by a tuple as:

$$MLDS = (U, C \cup L, V) \tag{3.1}$$

Let $A = C \cup L$, and $V = \cup_{a \in A} V_a$ where V is a union of the value domains. Each conditional feature $c \in C$ forms a subjective function $c : U \to V_c$, where V_c is the value domain of c and each label $l \in L$ forms a subjective function $l : U \to V_l$, where $V_l = \{0, 1\}$ is the value domain of l. If the object x is associated with label l, then $l(x) = 1$, otherwise, $l(x) = 0$.

Currently, some granular computing methods mainly deal with single-label data or multi-label data in rough set theory. Nevertheless, the data types in the multi-label learning are often complicated in practical applications. There are two types of data, i.e., continuous data and categorical data. Significant information is lost while processing continuous numerical data to be discreted. For this purpose, a neighborhood system is used for handling this problem.

Definition 2. For a multi-label decision system $MLDS = (U, C \cup L, V)$, given a certain threshold δ, the multi-label neighborhood decision system (MNDS) is defined by:

$$MNDS = (U,\ C \cup L,\ V,\ \delta) \tag{3.2}$$

where $\delta_B(x_i)$ is the δ-neighborhood set of instance x_i [61], which is defined as:

$$\delta_B(x_i) = \{x_j | x_j \in U, \Delta_B(x_i, x_j) \leq \delta\} \tag{3.3}$$

where Δ is a metric function and $\delta_B(x_i)$ is the information granule centered with sample x_i. The size of the neighborhood depends on the threshold δ.

Theorem 1. *Given a multi-label neighborhood decision system $MNDS = (U,\ C \cup L,\ V,\ \delta)$, let $A = C \cup L$, given any $P, Q \subseteq A$, for $\exists x \in U$, we have:*

1. *if $P \subseteq Q$, $\delta_Q(x) \subseteq \delta_P(x)$.*
2. *if $0 \leq \delta^1 \leq \delta^2 \leq 1$, $\delta_P^1(x) \subseteq \delta_P^2(x)$.*
3. *$\delta_Q(x) \subseteq \bigcap_{q \in Q} \delta_q(x)$.*
4. *$\delta_Q(x) \neq \emptyset$ and $\bigcup_{x \in U} \delta_Q(x) = U$.*

(1)

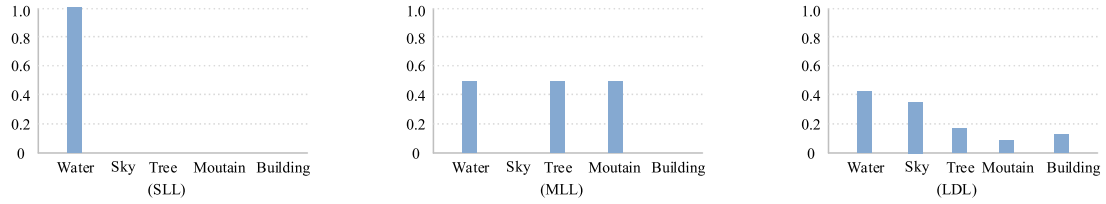


Fig. 1. One picture with different labels for three different ways.

Definition 3. Suppose $MNDS = (U,\ C \cup L,\ V,\ \delta)$ be a multi-label neighborhood decision system. Given any feature subset $B \subseteq C$ and $c \in B$, the B-lower and B-upper with the accuracy for any object subset $X \subseteq U$ is defined:

$$B_*(X)_\delta = \{x \in U | \delta_c(x) \subseteq X\} \tag{3.4}$$

$$B^*(X)_\delta = \{x \in U | \delta_c(x) \bigcap X \neq \emptyset\} \tag{3.5}$$

$$\alpha_B^\delta(X) = \frac{|B_*(X)_\delta|}{|B^*(X)_\delta|} \tag{3.6}$$

Lemma 1. *For a given multi-label neighborhood decision system $MNDS = (U,\ C \cup L,\ V,\ \delta)$, given two feature subset $C_1, C_2 \subseteq C$, for each instance subset $X \subseteq U$, if $C_1 \subseteq C_2$, then $\alpha_{C_1}^\delta(X) \leq \alpha_{C_2}^\delta(X)$.*

Proof. Since $C_1 \subseteq C_2 \subseteq C$, according to Theorem 1, we have $\delta_{C_2}(x) \subseteq \delta_{C_1}(x)$. From the Definition 3, we can obtain that $C_{1*}(X)_\delta \subseteq C_{2*}(X)_\delta$ and $C_1^*(X)_\delta \supseteq C_2^*(X)_\delta$. Consequently, $|\frac{C_{1*}(X)_\delta}{C_1^*(X)_\delta}| \leq |\frac{C_{2*}(X)_\delta}{C_2^*(X)_\delta}|$, subsequently $\alpha_{C_1}^\delta(X) \leq \alpha_{C_2}^\delta(X)$. □

Most of existing multi-label learning approaches assume the significance of different related label is equivalent. Therefore, we take the labeling-significance into account based on label distribution learning combined with granular computing. Nevertheless, the value of the labels in multi-label learning is either 1 or 0 rather than the distribution among [0, 1]. Therefore, in this section, some basic concepts on label distribution learning are introduced, and subsequently, the logical relationship is transformed into the label distribution. On this basis, a label enhancement algorithm based on the random variable distribution is proposed.

3.2. Label distribution decision system

In this subsection, we firstly reviewed the basic theories about label distribution learning [17,18,62]. Based on the multi-label neighborhood decision system, we define a novel label distribution decision system to deal with the problem in rough set.

For a training set $\Theta = \{(x_i, l_j) | 1 \leq i \leq n, 1 \leq j \leq m\}$, the project of multi-label learning is to evoke a multi-label predictor $\kappa : \mathbf{X} \Rightarrow 2^L$. Although the significant information is not explicitly accessible from the training samples, the significance of related labels existed in real-world multi-label learning problems is generally different. As Fig. 1 shown, *Picture(1)* describes various objects, such as the Water, Sky, Tree, Mountain and Building. In *Single-label learning,* the *Picture* may be labeled with the only object such as the Water. In *Multi-label learning,* it may be labeled with Water, Tree, Mountain. Both of the above ways only aim to answer the question "which object can describe the picture". However, in *Label distribution learning*, the *Picture* is labeled with the description degree of each object, which can deal with the problem "how much does each object describe the picture". Generally speaking, when we focus on the mapping from the samples to the labels, labeling one sample with a label distribution can be more efficient and accurate.

To denote the implicit relative label significance, $d_{x_i}^{y_j}$ is used for describing any sample $x_i \in X$ and the each label $y_j \in Y$. Correspondingly, we determine that the related label set Y for the sample x_i as $Y = \{y_j | d_{x_i}^{y_j} > \theta(x_i), 1 \leq j \leq m\}$, where the related labels are separated by the threshold $\theta(x_i)$ from the unrelated labels for the sample x_i. In order to effectively improve the precision of the learning algorithm, label distribution is used for traditional multi-label learning. Therefore, the basic definition for label distribution learning is stated. Given $\forall x_i \in X$, and each label $y_j \in Y$ $(1 \leq j \leq m)$ where m is denoted as the m-class labels, $d_{x_i}^{y_j}$ is the *Label-related Description Degree* [17,62].

Definition 4. Let U be the nonempty finite sample set. Let $C = \{c_1, c_2, \ldots, c_p\}$ be the feature set and $Y = \{y_1, y_2, \ldots, y_m\}$ be a label-related description degree set. Subsequently, a label distribution decision system (LDDS) can be defined by:

$$LDDS = (U, C \cup Y, V) \tag{3.7}$$

where V is with the same definitions in *MLDS*.

3.3. Label enhancement algorithm

In multi-label learning, the label significance is regarded as equivalent. However, in real-world applications, the description

of each label may be different [17,47]. In this paper, we introduce the distribution for multi-label classification. To convert the multiple label data into label distribution, we propose a label enhancement algorithm based on the random variable distribution in multi-label decision system. Given a label distribution decision system $LDDS = (U, C \cup Y, V)$ contained with a *N-dimension* samples space. Let $y = (y_1, y_2, \ldots, y_m)$ denote the label description degree associated with the sample x_i. Let X be a random variable that can take on the values $\{x_1, x_2, \ldots, x_n\}$ with the respective probabilities $\{p(x_1), p(x_2), \ldots, p(x_n)\}$ and $\sum_{i=1}^{n} p(x_i) = 1$.

Accordingly, we consider random variable distribution for the application to the label distribution learning. Based on the concepts of label distribution learning, the random variable distribution is satisfied the requirements of the probability statistics model. Additionally, it is known that neighborhood rough set model can generate similarity relationship among instances for multiple label. Therefore, the logical labels in multi-label learning can be converted into the label-distributed with its description degree using random variable distribution based on neighborhood rough set model.

Definition 5. Given a multi-label neighborhood decision system $MLNS = (U,\ C \cup L,\ V,\ \delta)$. Let $L = \{l_1, l_2, \ldots, l_m\}$ be the label set. For any sample x with its neighborhood $\delta(x) = \{x, \ldots, x_k, \ldots, x_n\}$, the label-related description degree for each label is defined as:

$$\forall x \in \delta(x),\ d_x^y = \frac{S_l[\delta(x)]}{S_L[\delta(x)]} \wedge (V_x^l = 1) \tag{3.8}$$

d_x^y is denoted as the label significance of label y for instance x where y is the enhanced label l. $V_x^l = 1$ represents that the value of label l for instance x is 1. $S_L[\delta(x_i)]$ is represented the summary of instance x with its neighborhood for label set L where $\forall l \in L, V_x^l = 1$. In addition, $S_l[\delta(x)]$ represents the summary of instance x with its neighborhood for label l where $V_x^l = 1$. Particularly, $d_x^y = 0$ where $V_x^l = 0$, and $d_x^{y_j} = \frac{d_x^{y_j}}{\sum_{j=1}^{m} d_x^{y_j}}$ where $\sum_{j=1}^{m} d_x^{y_j} < 1$.

Theorem 2. *For each label significance of label y_j for instance x, the terms $d_x^{y_j} (1 \le j \le m)$ should satisfy the two conditions as follows:*

1. *normalization constraint:* $\sum_{j=1}^{m} d_x^{y_j} = 1$.
2. *non-negativity constraint:* $d_x^{y_j} \ge 0$.

Lemma 2. *Suppose $MLNS = (U,\ C \cup L,\ V,\ \delta)$ be a multi-label neighborhood decision system. Given two non-negative neighborhood threshold δ^1 and δ^2. For each logical label l, the enhanced labeling-significance d_x^y. If $\delta^1 \le \delta^2$, we have $\forall l \in L$, $d_x^{y^{(1)}} \le d_x^{y^{(2)}}$.*

Proof. As two given neighborhood threshold $\delta^1 \le \delta^2$ for each label l, according to Theorem 1, we have $\delta_l^1(x) \subseteq \delta_l^2(x)$, and we can obtain $S_l[\delta_l^1(x)] \le S_l[\delta_l^2(x)]$. Subsequently, according to Definition 5, we can obtain $d_x^{y^{(1)}} = \frac{S_l[\delta_l^1(x)]}{S_L[\delta_L(x)]}$ and $d_x^{y^{(2)}} = \frac{S_l[\delta_l^2(x)]}{S_L[\delta_L(x)]}$, and then we can obtain $d_x^{y^{(1)}} - d_x^{y^{(2)}} = \frac{S_l[\delta_l^1(x)] - S_l[\delta_l^2(x)]}{S_L[\delta_L(x)]} \le 0$. Consequently, $d_x^{y^{(1)}} \le d_x^{y^{(2)}}$. □

From aforementioned analysis, considering labeling-significance, i.e., significance of related labels, is necessary for multi-label learning. Nevertheless, the exiting multi-label data assume the labels perform with uniform distribution with logical labels. Subsequently, in order to convert each multiple label into the label distribution with more supervised information, a label enhancement algorithm using the random variable distribution and granular computing is given in Algorithm 1. The crucial procedure is to calculate the possibility of the positive described label associated with the neighborhood of each instance in the algorithm. In practical applications, the threshold of neighborhood granularity (δ) is significant for label enhancement. The neighborhood may not contain any instance when δ is significantly small, though it cannot reflect the information while increasing the size of the neighborhood. In this paper, the Euclidean distance is used for calculating the neighborhood granularity. The threshold δ of the neighborhood is computed by $\delta = \frac{1}{\theta}\mathbb{M}[\sum_{k=1}^{n} \frac{\sigma(a_k)}{\theta}]$ where $\sigma(a_k)$ is represented as the standard deviation of the feature a_k, and $\mathbb{M}[\bullet]$ is denoted as the average of the standard deviation for the whole feature set. The θ is described in detail and it can be obtained in [61].

The neighborhood of each sample x should be computed firstly in Step 2. Some initialization are given in Step 3. S_x^L is denoted as the summary of the whole label set L valued 1 for sample x, and S_x^l is denoted as the summary of label l valued 1 for sample x. From Step 5 to Step 11, the key step is to compute the S_x^L and S_x^l for each sample x. Based on the above, the label-related description degree can be obtained by Definition 4 in Step 6 and Step 7. The Step 8 to 10 is to ensure the $Sum(d_x^Y)$ in Step 10 satisfy the condition $\sum_{j=1}^{m} d_x^{y_j} = 1$.

Example 1. Given a multi-label neighborhood decision system $MNDS = (U,\ C \cup L,\ V,\ \delta)$, and the label set $L = \{l_1, l_2, l_3, l_4\}$. Given an instance x_1 with labels $L(x_1) = \{1, 1, 0, 1\}$. Assuming that $\delta_C(x_1) = \{x_1, x_5\}$ and $L(x_5) = \{0, 1, 1, 0\}$. On this basis, we can compute the frequency of each label $S_{l_1}[\delta(x_1)] = 1,\ S_{l_2}[\delta(x_1)] = 2,\ S_{l_3}[\delta(x_1)] = 1,\ S_{l_4}[\delta(x_1)] = 1$. Therefore, we can obtain the $S_L[\delta(x_1)] = 5$. Subsequently, through Definition 4 and Algorithm 1, we can obtain $d_{x_1}^{y_j}$ for each label $d_{x_1}^{y_1} = \frac{1}{5},\ d_{x_1}^{y_2} = \frac{2}{5},\ d_{x_1}^{y_3} = 0,\ d_{x_1}^{y_4} = \frac{1}{5}$. However, $Sum(d_{x_1}^Y) = \frac{4}{5} \ne 1$, and we should transform $d_{x_1}^{y_j}$ for each label, therefore $d_{x_1}^{y_1} = \frac{1}{5}/\frac{4}{5} = \frac{1}{4},\ d_{x_1}^{y_2} = \frac{2}{5}/\frac{4}{5} = \frac{1}{2},\ d_{x_1}^{y_3} = 0,\ d_{x_1}^{y_4} = \frac{1}{5}/\frac{4}{5} = \frac{1}{4}$.

4. Label distribution feature selection based on mutual information

Various of feature selection methods are proposed to find relevant and discriminating features to determine the output labels. In this section, the main concepts for feature selection are discussed [15,31] and a new feature selection method based on mutual information for label distribution is proposed. Shannon's entropy [63,64] is a significant measure which is used to investigate the uncertainty of random variables.

Definition 6. Given a set $S = \{s_1,\ s_2,\ \cdots,\ s_m\}$, the entropy of random variable S is defined as:

$$H(S) = -\sum_{i=1}^{m} p(s_i) \log\ p(s_i) \tag{4.1}$$

where the $p(s_i)$ is denoted as the probability of s_i. Given a new random variable set $T = \{t_1,\ t_2,\ \cdots,\ t_n\}$ and $p(t_j)$ is denoted as the probability of t_j. If the variable S is known where the entropy of T is after observing the values of S, there is a definition namely the conditional entropy for measuring the remaining uncertainty where $p(t_j \mid s_i)$ is the conditional probability:

$$H(T \mid S) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(s_i,\ t_j) \log\ p(t_j \mid s_i) \tag{4.2}$$

where the $p(s_i,\ t_j)$ is denoted as the joint probability between S and T. Accordingly, the joint entropy is defined as follow:

$$H(S,\ T) = H(S) + H(T \mid S) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(s_i,\ t_j) \log\ p(s_i,\ t_j) \tag{4.3}$$

Algorithm 1 A Label Enhancement algorithm for Multi-label Neighborhood Decision System (LEMN)

Input: A multi-label neighborhood decision system $MNDS = (U,\ C \cup L,\ V,\ \delta)$.

Output: A label distribution decision system $LDDS = (U,\ C \cup Y,\ V)$.

Begin

1. *Let $x \in X$, $l \in L$, $y \in Y = \emptyset$;*
2. *Compute neighborhood $\delta(x)$ for each sample x;*
3. *Let $S_l[\delta(x)] = 0$, $S_L[\delta(x)] = 0$; // $S_L[\delta(x)]$ is the summary of label set L valued 1, $S_l[\delta(x)]$ is the summary of label l valued 1*
4. **Repeat**
5. *Compute $S_l[\delta(x)]$, $S_L[\delta(x)]$ for each sample x;*
6. *Compute the significance d_x^y of label y by Formula 3.8;*
7. *Let $Sum(d_x^Y) = 0$; // $Sum(d_x^Y)$ is the summary of each label's significance of Y*
8. *Compute $Sum(d_x^Y)$;*
9. *if $Sum(d_x^Y) < 1$: Transform d_x^y by the Definition 5;*
10. **Until** $x > n$

End

Lemma 3. *Given a multi-label neighborhood decision system $MNDS = (U,\ C \cup L,\ V,\ \delta)$. For $B_1 \subseteq B_2 \subseteq C$, we have $H_\delta(B_1) \leq H_\delta(B_2)$.*

Proof. As $x \in U$, we have $\delta_{B_2} \subseteq \delta_{B_1}$ where $B_1 \subseteq B_2$. Let $[x]_L$ be the equivalence relation. According to Theorem 1, we have $\{x\} \subseteq (\delta_{B_2}(x) \cap [x]_L) \subseteq (\delta_{B_1}(x) \cap [x]_L) \subseteq U$. Then, $\frac{1}{|U|} \leq \frac{|(\delta_{B_2}(x) \cap [x]_L)|}{|U|} \leq \frac{|(\delta_{B_1}(x) \cap [x]_L)|}{|U|} \leq 1$ and $log\,(\frac{1}{|U|}) \leq log\,(\frac{|(\delta_{B_2}(x) \cap [x]_L)|}{|U|}) \leq log\,(\frac{|(\delta_{B_1}(x) \cap [x]_L)|}{|U|}) \leq 0$. Thus according to Definition 2, $0 \leq -\frac{1}{|U|}\sum_{i=1}^{|U|} log\,(\frac{|(\delta_{B_1}(x) \cap [x]_L)|}{|U|}) \leq -\frac{1}{|U|}\sum_{i=1}^{|U|} log\,(\frac{|(\delta_{B_2}(x) \cap [x]_L)|}{|U|}) \leq log\,(|U|)$. Consequently, $H_\delta(B_1) \leq H_\delta(B_2)$. □

Mutual information is widely used for measuring the dependency of variables. Some basic concepts on mutual information will be given [31].

To measure the dependency of S, which decreased after observing the variable set T, the mutual information is defined as follow, which can quantify the information-acquired of S from T:

$$I(S;\ T) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(s_i,\ t_j) \log \frac{p(s_i,\ t_j)}{p(s_i)p(t_j)} \tag{4.4}$$

where $p(s_i)$ is denoted as the probability of s_i and $p(t_j)$ is denoted as the probability of t_j.

Lemma 4. *Mutual information can be denoted by the related individual entropy as follow:*

$$I(S;\ T) = H(S) + H(T) - H(S,\ T) \tag{4.5}$$

Proof. $I(S;\ T) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(s_i,\ t_j) \log \frac{p(s_i,\ t_j)}{p(s_i)p(t_j)}$

$= \sum_{i=1}^{m}\sum_{j=1}^{n} p(s_i,\ t_j) \log\ p(s_i,\ t_j) - \sum_{i=1}^{m}\sum_{j=1}^{n} p(s_i,\ t_j) \log\ p(s_i)p(t_j)$

$= \sum_{i=1}^{m}\sum_{j=1}^{n} p(s_i,\ t_j) \log\ p(s_i,\ t_j) - \sum_{i=1}^{m}\sum_{j=1}^{n} p(s_i,\ t_j) \log\ p(t_j)$

$= \sum_{i=1}^{m}\sum_{j=1}^{n} p(s_i,\ t_j) \log\ p(s_i,\ t_j) - \sum_{i=1}^{m} p(s_i) \log\ p(s_i) - \sum_{j=1}^{n} p(t_j) \log\ p(t_j)$

$= H(S) + H(T) - H(S,\ T)$ □

It is known that feature selection has effectiveness on keeping the original feature space differentiated. In this section, the dependency between features and labels is introduced and the algorithm for label distribution feature selection using mutual information is proposed.

Definition 7. Given one random feature $c \in C$ and a set of enhanced-labels $Y = \{y_1, y_2, \ldots, y_m\}$. The mutual information between c and Y is defined as:

$$MI(Y;\ c) = \sum_{j=1}^{m} I(y_j;\ c) \tag{4.6}$$

According to information theory, the following are the properties of the mutual information between features and the set of labels.

1. *non-negativity:* $MI(Y,\ c) \geq 0$.
2. *normalization:* $MI(Y,\ c) \leq \sum_{j=1}^{m} H(y_j)$.

The mutual information between feature c and label set Y is the minimum, where c is independent of each label y_j, and otherwise it is the maximum.

Definition 8. Given a threshold Θ and the mutual information between features and labels $MI(Y,\ c)$, the relationship between Θ and $MI(Y,\ c)$ is given:

1. *Dependency:* $|MI(Y,\ c)| \geq \Theta$.
2. *in-dependency:* $|MI(Y,\ c)| < \Theta$.

To normalize the values in [0, 1] range with value 1 indicating complete dependency and value 0 indicating complete independence of Y and c, where the mutual information gain is biased in favor of features with more values. A filter method based on the concept of entropy in information theory is employed to measure correlation between features and labels and it is defined as:

$$NMI(Y,\ c) = 2 \times \frac{MI(Y,\ c)}{H(Y) + H(c)} \tag{4.7}$$

Therefore, $NMI(y_j,\ c) \in [0,\ 1]$. It is independent for variables y_j and c where $NMI(y_j,\ c) = 0$. On the other hand, it is indicated that y_j and c are dependent, where $NMI(y_j,\ c) = 1$. In addition, $NMI(Y,\ c)$ can be transformed into the *mutual information based on Standard Gaussian distribution* $SMI(Y,\ c)$ through the formula:

$$SMI(Y,\ c) = \frac{NMI(Y,\ c) - \mu}{\theta} \tag{4.8}$$

where θ is denoted as the *standard deviation* of the mutual information between features and labels, and μ is the mean.

It is assumed that the number of candidate labels is $k = 10$, and given two different features c_1 and c_2 when calculating the mutual information between the features and the labels. For

example, assuming that $SMI(Y,\ c_1) = 2.8$ and $SMI(Y,\ c_2) = 1.75$, accordingly, feature c_1 will be reserved because the significance of c_1 is greater than c_2. However, it has probability $MI(y_1,\ c_1) = MI(y_2,\ c_1) = \cdots = MI(y_{10},\ c_1) = 0.28$, whereas for c_2, $MI(y_1,\ c_2) = 0.75$, $MI(y_2,\ c_2) = 1$ and $MI(y_j,\ c_2) = 0$ for $j \in [3,\ 10]$. Conditionally, the recognition ability of c_2 is much larger than the feature c_1; hence, c_2 should be reserved. In order to deduce the occurrence of the situation, the standard distribution threshold after the above is performed can be obtained:

$$\Theta_{sd} = \frac{1}{k}\sum_{i=1}^{k} |SMI(Y,\ c_i)| \tag{4.9}$$

According to Definition 8, a feature significance measurement is given.

Definition 9. Given a standard distribution threshold Θ_{sd}, let $C = [c_1, c_2, \ldots, c_p]$ be the set of features and $Y = [y_1, y_2, \ldots, y_m]$ be the set of enhanced-labels in label distribution, whether the feature $c \in C$ is associated with the labels is defined as:

1. *dependency of feature-label*: $|SMI(Y,\ c)| \geq \Theta_{sd}$
2. *in-dependency of feature-label*: $|SMI(Y,\ c)| < \Theta_{sd}$

Multi-label data with conventional logical labels have been converted into label distribution data by using the label enhancement algorithm. To improve the classification performance for multi-label data, a feature selection algorithm based on mutual information embedded in label distribution is developed. The framework of the proposed algorithms is shown in Fig. 2.

On this basis, to remove some redundant or irrelevant features in multi-label data, we design a label distribution feature selection algorithm based on mutual information in Algorithm 2. Computing the *mutual information based on Standard Gaussian distribution (SMI)* between each feature and the label set is the primary process in this algorithm.

Firstly, assuming that the universe be $|U|$, the feature set is $|C|$, the feature subset is $|S|$ and the label set is $|Y|$. Based on Definition 7 and Formula (4.7), *mutual information (MI)* and the *mutual information normalized* (*NMI*) can be computed from Step 2 to Step 6, and the time complexity is $O(|U|C|Y|)$ when computing the equivalent relation based on the radix sorting. Then, according to the Formula (4.7), we can convert the *NMI* of each label into the *SMI* from Step 8 to Step 11, and the time complexity is $O(|C||Y|)$. The threshold Θ_{sd} can be calculated by Formula (4.9) in Step 12. Finally, the feature related to the labels can be selected by the threshold Θ_{sd} from Step 13 to Step 16 and the time complexity is $O(|C|)$. The selected and sorted-descending feature subset by *SMI* is outputted in Step 17, and the time complexity is $O(|S|log|S|)$.

5. Experimental analysis

In this section, experiments are conducted to evaluate the performance of the proposed feature-selection algorithm.

5.1. Data sets

In order to validate the performance of the proposed method, 12 real world multi-label data sets are selected from the Mulan library[1] and the multi-label learning Resources,[2] namely *Flags*, *Emotions*, *CAL500*, *Yeast*, *Birds*, *Scene*, *3sources(bbc1000)*, *Plant(PseAAC)*, *WaterQuality*, *Gnegative(PseAAC)*, *Human(PseAAC)*, *Eukaryote(PseAAC)*, respectively. The 12 data sets are selected from six distinct practical application domains such as audio, music, image, biology and text. The detailed description of them is displayed in Table 1.

Table 1
12 multi-label data-sets used for experiments.

No.	Data set	Domain	Samples	Features	Labels
1	Birds	audio	645	260	19
2	Emotions	music	593	72	6
3	CAL500	music	502	68	174
4	Flags	image	194	19	7
5	Scene	image	2407	294	6
6	Yeast	biology	2417	103	14
7	Plant	biology	978	440	12
8	WaterQuality	chemistry	1060	16	14
9	Gnegative	biology	1392	440	8
10	3sources	text	352	1000	6
11	Human	biology	3106	440	14
12	Eukaryote	biology	7766	440	22

5.2. Multi-label learning classification performance

To evaluate the multi-label learning classification performance, six evaluation metrics [9,53], namely *Hamming Loss*, *Ranking Loss*, *Coverage*, *One Error*, *Average Precision* and *Subset Accuracy*, are utilized to testify the performance of the proposed algorithms. Assuming that $d_i \subseteq D$ is the vector of the labels. The description of each performance is noted:

- Hamming Loss (*HL*): it evaluates that how many times an sample-label pair is misclassified and the formula is defined as follow:

$$HL = \frac{1}{n}\sum_{i=1}^{N}\frac{|D_i' \oplus d_i|}{m}$$

  where $\oplus$ is denoted as the *XOR* operation and the smaller value, the better performance, which concerns the label prediction.
- Ranking Loss (*RL*): it evaluates the fraction of reversely ordered label pairs and it is expressed as follow:

$$RL = \frac{1}{n}\sum_{i=1}^{N}\frac{|1}{|d_i||\overline{d_i}|}\ |\{(k_1,\ k_2)|k_1 \leq k_2, (k_1,\ k_2) \in d_i \times \overline{d_i}\}$$

  where k_i is a real-valued likelihood between x_i and each label $y_i \in L$ based on a multi-label classifier, and $\overline{d_i}$ denotes the complementary set of d_i. It concerns with the label ranking, and the smaller value, the better performance.
- Coverage (*CV*): it evaluates the average distance which is supposed to be dropped the label ranking list to cover all the ground truth labels of the sample with the definition as follow:

$$CV = \frac{1}{n}\sum_{i=1}^{N}\max_{j\in d_i} rank(j) - 1$$

  where $rank(j)$ is denoted as the rank list of j according to its likelihood. Accordingly, the $rank(k_1) < rank(k_2)$ where $k_1 < k_2$. It concerns the label ranking, and the smaller value, the better performance.
- One Error(*OE*): it evaluates the fraction of samples with top-ranked label which is not the vector of proper labels, and it is expressed as follow:

$$OE = \frac{1}{n}\sum_{i=1}^{N}[[\underset{d_i\in D}{argmax} f(x_i, d_i)] \notin D_i']$$

[1] http://mulan.sourceforge.net/datasets-mlc.html
[2] http://www.uco.es/kdis/mllresources

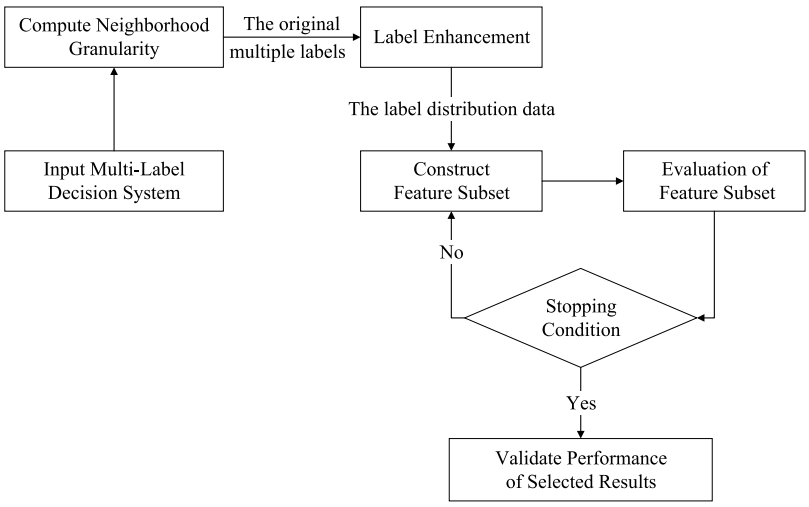


Fig. 2. The process of proposed algorithms.

Algorithm 2 Label Distribution feature selection based on Mutual Information (LDMI)

Input: A label-distribution decision system $LDDS = (U, C \cup Y, V)$.
Output: A feature-selected subset S.
Begin
1. *Let* $S = \emptyset$, $Label_Y = \emptyset$; // The $Label_Y$ is used for saving the results of *NMI* for each label
2. ***For*** $c_i \in C, y_j \in Y$:
3. *compute mutual information* $MI(y_j, c_i)$;
4. *let the normalized mutual information* $NMI(y_j, c_i))) \leftarrow MI(y_j, c_i)))$ *by Formula. 4.7;*
5. *let* $Label_Y \leftarrow Label_Y \cup NMI(y_j, c_i)$;
6. ***End For***
7. *Let the mutual information based on Standard Gaussian distribution* $SMI = \emptyset$;
8. ***For*** each NMI_c in $Label_Y$:
9. *let* $SMI(Y,\ c))) \leftarrow NMI(Y,\ c)))$ *by Formula. 4.8;*
10. *let* $SMI \leftarrow SMI \cup SMI(Y,\ c)))$;
11. ***End For***
12. *Compute the threshold* Θ_{sd} *by Formula. 4.9;*
13. ***For*** each SMI_c in SMI :
14. if $|SMI_c| \geq \Theta_{sd}$
15. *let* $S \leftarrow S \cup \{c\}$;
16. ***End For***
17. *Sorted*(S) order by $|SMI(Y,\ c)|$;
18. ***Return*** S
End

if any predicate ε holds, then $[[\varepsilon]]$ equals 1 and 0 otherwise. And it concerns the label ranking, the smaller the better.

- Average Precision (AP): it evaluates the ranked label's average precision for the label $d \in d_k$ with the definition:

$$AP = \frac{1}{n}\sum_{i=1}^{N}\frac{1}{|d_i|}\sum_{d\in d_i}\frac{|\{d' \in d_i : k_i(d') \leq k_i(d)\}|}{k_i(d)}$$

where $k_i(d)$ is denoted as the ranked label $d \in D$ predicted respectively, and it concerns with the label ranking, the larger the better.

- Subset Accuracy(SA): it calculates the ratio of test samples correctly to the carnality of test set and it is defined as:

$$SA = \frac{1}{n}\sum_{i=1}^{n}[Z_i = d_i]$$

where $Z_i \subseteq D$ is denoted as a predicted label set which associated with the sample x_i.

5.3. Experimental design and settings

In this section, to evaluate the performance of our proposed algorithm persuasively and validate the efficiency, we conduct experiments to compare the performances with other state-of-art multi-label feature selection methods, including *PMU*[65], *MDDMproj* [66], *MDDMspc* [66], *RF-ML* [67] and *MLFRS* [21]. To make comparisons more effective and reliable, the parameter settings are optimized supplied by the recommendation of all the comparing algorithms. All of the detailed information are shown and the basic parameters from the methods are set as follows:

- PMU: A multi-label classification feature selection method using mutual information that is applied to several multi-label classification problems. The continuous features are discretized into 4 bins by equal-width strategy, and the others are as recommended in [65].
- MDDM: A method via dependence maximization that tries to identify a lower-dimensional feature space maximizing the dependence between the original feature description and class labels related to the object. In addition, it includes

(a) Hamming Loss(↓) (b) One Error(↓)

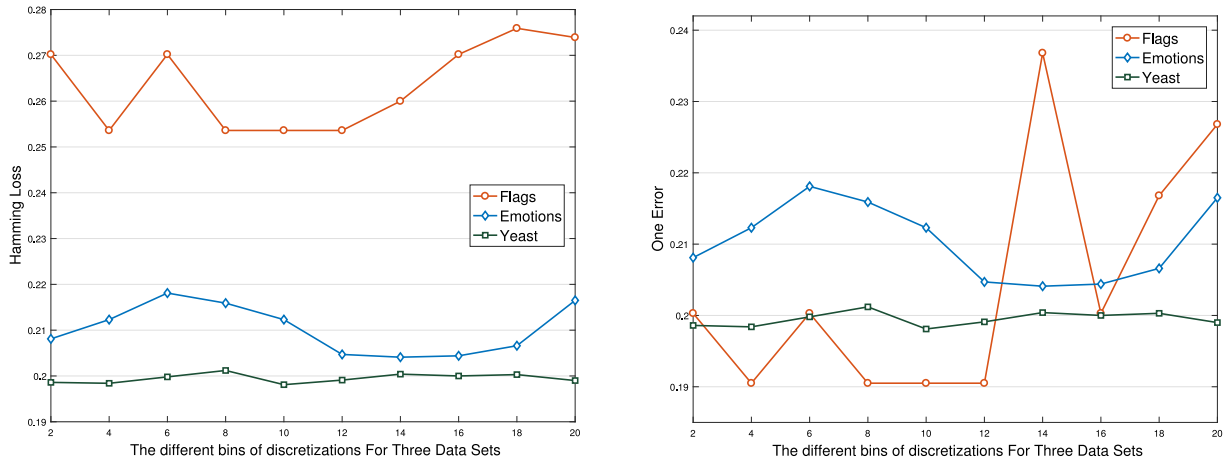
(c) Coverage(↓) (d) Ranking Loss(↓)

(e) Average Precision(↑) (f) Subset Accuracy(↑)

Fig. 3. Six evaluation metrics by different bins of discretization on three data sets from DMLkNN.

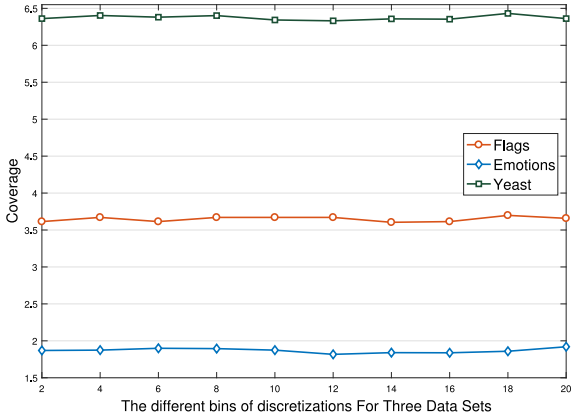
the $MDDM_{proj}$ and $MDDM_{spc}$. The parameter μ is as 1.0 as suggested in [66].

- *RF-ML*: A new multi-label feature selection algorithm by extending the single-label feature selection *ReliefF* algorithm [67] that takes the effectiveness of interacting attributes into account to deal with multi-label data without any data transformation directly**.** For experiments, the number of the *k-nearest* neighbors is set as 5 as suggested in [67].
- *MLFRS*: This method used fuzzy relations to measure the similarity between instances under different labels, and selected the optimal multi-label feature subset based on a forward greedy feature selection algorithm. For experiment,

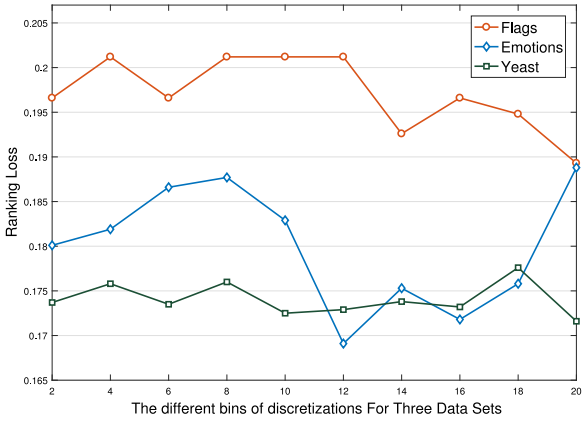


(a) Hamming Loss(↓)

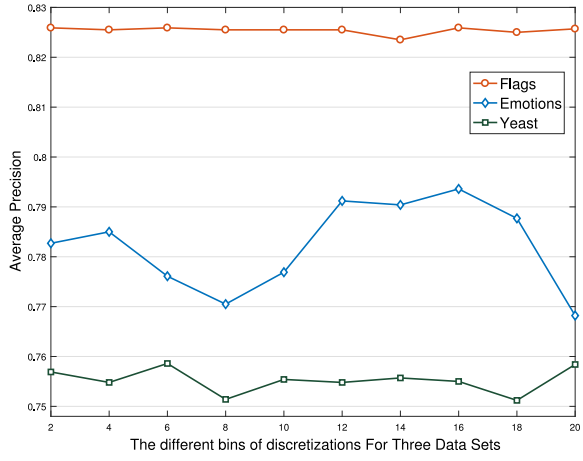
(b) One Error(↓)



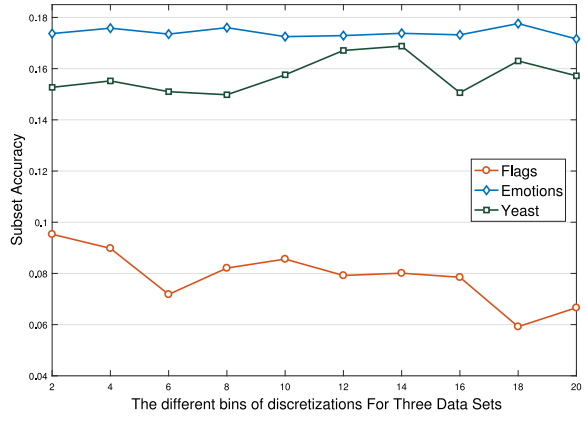
(c) Coverage(↓)



(d) Ranking Loss(↓)



(e) Average Precision(↑)



(f) Subset Accuracy(↑)

Fig. 4. Six evaluation metrics by different bins of discretization on three data sets from MLkNN.

the sampling rate of sample section is set to {0.15, 0.1, 0.05} as suggested in [21].

For our proposed algorithm, the continuous variables from the data sets are discretized from 2 bins to 20 bins, and the 10-fold invalidation method is used for the over-fitting problem. In multi-label classification learning, there are six classification performance metrics namely *One-error, Hamming Loss, Coverage, Ranking Loss, Average precision* and *Subset Accuracy*. The classification performance of all feature selection algorithms are evaluated by two multi-label classifiers *MLkNN* and *DMLkNN* in this paper.

5.4. Experimental results and analysis

In this section, we perform the experimental results by comparing the six multi-label feature selection algorithms. First, the

Table 2
Hamming Loss (↓) of six feature selection algorithms for DMLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.0506(6)	0.0490(3)	0.0471(2)	0.0505(5)	0.0498(4)	**0.0468(1)**
Flags	0.2786(3)	0.2891(6)	0.2829(5)	**0.2745(1.5)**	0.2822(4)	**0.2745(1.5)**
Emotions	0.2010(2)	0.2019(4)	0.2013(3)	0.2024(5)	0.2061(6)	**0.2004(1)**
CAL500	0.1380(5)	0.1380(5)	0.1380(5)	0.1376(3)	**0.1375(1.5)**	**0.1375(1.5)**
Yeast	0.1994(5)	0.1991(4)	0.2015(6)	0.1985(2)	0.1988(3)	**0.1984(1)**
Scene	0.0965(2)	0.0974(3)	0.1256(5)	0.1332(6)	0.1127(4)	**0.0933(1)**
Plant	0.0844(5.5)	**0.0814(1)**	0.0837(4)	0.0824(3)	0.0844(5.5)	0.0820(2)
WaterQuality	0.3045(2.5)	0.3057(4)	0.3072(5)	0.3266(6)	0.3045(2.5)	**0.3024(1)**
Gnegative	0.0589(2)	0.0730(5)	0.0699(4)	0.0794(6)	0.0652(3)	**0.0545(1)**
3sources	0.1871(4)	**0.1866(1.5)**	0.1871(4)	0.1876(6)	0.1871(4)	**0.1866(1.5)**
Human	0.0819(5)	0.0801(2)	0.0802(3)	0.0816(4)	0.0820(6)	**0.0800(1)**
Eukaryote	**0.0510(1.5)**	0.0511(3)	0.0514(4)	0.0516(6)	0.0515(5)	**0.0510(1.5)**
Avg. rank	3.62	3.45	4.17	4.46	4.04	**1.25**

Table 3
One Error (↓) of six feature selection algorithms for DMLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.7283(5)	0.7282(3)	0.7236(4)	0.7469(6)	0.7189(2)	**0.7080(1)**
Flags	0.1800(2)	0.1961(4)	0.2105(6)	0.1897(3)	0.2005(5)	**0.1747(1)**
Emotions	0.2858(6)	**0.2649(1.5)**	0.2800(4)	0.2969(5)	0.2717(3)	**0.2649(1.5)**
CAL500	0.1157(3.5)	0.1157(3.5)	0.1175(5)	0.1176(6)	0.1156(1.5)	**0.1156(1.5)**
Yeast	0.2358(4)	0.2379(5)	0.2416(6)	0.2251(3)	0.2238(2)	**0.2222(1)**
Scene	0.2613(2)	0.2659(3)	0.3149(5)	0.3722(6)	0.3012(4)	**0.2505(1)**
Plant	0.6742(5.5)	0.6599(2.5)	0.6742(5.5)	0.6701(4)	0.6599(2.5)	**0.6579(1)**
WaterQuality	0.2977(3)	0.3009(4)	0.2972(2)	0.3462(6)	0.3113(5)	**0.2934(1)**
Gnegative	0.6583(4)	0.6411(3)	0.6145(2)	0.6677(6)	0.6598(5)	**0.5995(1)**
3sources	**0.6873(1)**	0.6884(4)	0.6897(5)	0.7106(6)	0.6878(3)	0.6876(2)
Human	0.6363(4)	0.6035(3)	0.6559(6)	0.6456(5)	0.5957(2)	**0.5938(1)**
Eukaryote	0.6507(4)	0.6364(3)	0.6212(2)	0.6555(5)	0.6647(6)	**0.6184(1)**
Avg. rank	3.67	3.29	4.3	5.0	3.42	**1.17**

Table 4
Coverage (↓) of six feature selection algorithms for DMLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	2.8532(4)	2.7770(3)	2.8915(5)	2.9704(6)	2.3600(2)	**2.3146(1)**
Flags	3.7068(3)	3.7587(5)	3.7358(4)	3.7653(6)	3.7005(2)	**3.6879(1)**
Emotions	1.8385(4)	1.8303(2)	1.8366(3)	1.8960(6)	1.8506(5)	**1.7647(1)**
CAL500	**128.9147(1.5)**	129.6664(5)	129.6564(4)	129.7004(6)	129.4623(3)	**128.9147(1.5)**
Yeast	6.3512(3)	6.4199(6)	6.3657(4)	6.4033(5)	**6.3450(1)**	6.3470(2)
Scene	**0.5164(1)**	0.5422(3)	0.7179(5)	0.7948(6)	0.6910(4)	0.5280(2)
Plant	2.0462(3)	2.0410(2)	2.1770(6)	2.1116(4)	2.1298(5)	**2.0001(1)**
WaterQuality	9.0491(4)	9.2151(5)	9.0434(3)	9.3670(6)	9.0132(2)	**9.0019(1)**
Gnegative	0.5901(4)	0.5462(3)	0.5148(2)	0.6677(6)	0.5952(5)	**0.4732(1)**
3sources	2.1389(2)	2.1801(3)	2.1995(6)	2.1811(4)	2.1883(5)	**2.1257(1)**
Human	2.3969(4)	2.2778(3)	2.4710(6)	2.3992(5)	2.1864(2)	**2.1745(1)**
Eukaryote	3.0264(4.5)	2.8324(2)	2.8886(3)	3.0264(4.5)	3.0572(6)	**2.7901(1)**
Avg. rank	3.16	3.50	4.07	5.38	3.61	1.11

effectiveness of the different discretization bins is analyzed. We take three data sets namely *Flags*, *Emotions* and *Yeast* varying from sizes small to large to evaluate the performance of different discretization bins for our proposed method.

As displayed in Figs. 3 and 4, a large difference among the different degree of discretization is obtained for *Flags*, which is close to 5% between the highest degree and the lowest degree. In addition, *Flags* obtains a better performance from 4 bins to 8 bins in *Average Precision* and *Subset Accuracy* and . For the medium data set *Emotions,* it showed a relatively stable performance on *Hamming Loss* and *Coverage*. On the other evaluation metrics, the discretizations bins from 12 bins to 16 bins outperforms the other bins for the proposed method. For the large data set *Yeast*, it showed a relatively stable performance with a small fluctuation for all the evaluation metrics. To sum up, the discretization bins may result in different evaluation performances on small data sets. For the medium and large data sets, the proposed algorithm obtains stable performance under the different discretization bins.

5.4.1. Evaluation of the classification performance

To demonstrate the effectiveness of the proposed algorithm *LDMI*, we make a comparison to predict the classification performance among *LDMI, MDDMspc, MDDMproj, PMU, RF-ML* and *MLFRS. LDMI* can obtain the feature subset directly, whereas *MDDMproj*, *MDDMspc, PMU*, *RF-ML* and *MLFRS* get the feature

Table 5
Ranking Loss (↓) of six feature selection algorithms for DMLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.1029(4)	0.0997(3)	0.1034(5)	0.1125(6)	0.0828(2)	**0.0788(1)**
Flags	0.2146(5)	0.2127(4)	0.2072(2)	0.2212(6)	0.2100(3)	**0.2035(1)**
Emotions	0.1747(4)	**0.1711(1)**	0.1732(3)	0.1755(6)	0.1752(5)	0.1727(2)
CAL500	0.1813(4)	0.1808(2.5)	0.1818(5.5)	0.1818(5.5)	0.1808(2.5)	**0.1803(1)**
Yeast	0.1749(5)	0.1766(6)	0.1725(4)	0.1718(3)	**0.1711(1)**	0.1717(2)
Scene	0.0861(2)	0.0912(3)	0.1266(5)	0.1420(6)	0.1210(4)	**0.0882(1)**
Plant	0.1883(6)	0.1745(2)	0.1757(3)	0.1824(4)	0.1839(5)	**0.1725(1)**
WaterQuality	0.2751(4)	0.2729(3)	0.2708(2)	0.3047(6)	0.2844(5)	**0.2675(1)**
Gnegative	0.0672(2)	0.0720(3)	0.0782(5)	0.0778(4)	0.0822(6)	**0.0617(1)**
3sources	0.3850(2)	0.4070(4)	0.4173(6)	0.4084(5)	0.4003(3)	**0.3804(1)**
Human	0.1630(4)	0.1500(3)	0.1686(6)	0.1633(5)	0.1455(2)	**0.1454(1)**
Eukaryote	0.1212(2)	0.1239(3)	0.1307(5)	0.1302(4)	0.1319(6)	**0.1204(1)**
Avg. rank	3	3.63	4.2	5.04	3.38	**1.15**

Table 6
Average Precision (↑) of Five feature selection algorithms for DMLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.5662(4)	0.5727(3)	0.5612(5)	0.5267(6)	0.5964(2)	**0.6198(1)**
Flags	0.8098(5)	**0.8188(1)**	0.8185(2)	0.8017(6)	0.8121(4)	0.8163(3)
Emotions	0.7905(3)	0.7913(2)	0.7870(6)	0.7898(4)	0.7893(5)	**0.7928(1)**
CAL500	0.4959(4)	0.4970(2)	0.4952(5)	0.4936(6)	0.4961(3)	**0.4979(1)**
Yeast	0.7542(5)	0.7559(4)	0.7493(6)	0.7580(2.5)	0.7580(2.5)	**0.7585(1)**
Scene	0.8462(2)	0.8406(3)	0.7785(5)	0.7686(6)	0.8066(4)	**0.8502(1)**
Plant	0.5450(5)	0.5588(2)	0.5587(3)	0.5461(4)	0.5449(6)	**0.5691(1)**
WaterQuality	0.6756(4)	0.6763(3)	0.6800(2)	0.6387(6)	0.6657(5)	**0.6830(1)**
Gnegative	0.7236(4)	0.7389(3)	0.7635(2)	0.7092(6)	0.7201(5)	**0.7791(1)**
3sources	0.5153(3)	0.5151(4)	0.5055(5)	0.5004(6)	0.5201(2)	**0.5224(1)**
Human	0.5571(4)	0.5803(3)	0.5914(2)	0.5532(6)	0.5436(5)	**0.5953(1)**
Eukaryote	0.5326(4)	0.5473(3)	**0.5664(1)**	0.5306(5)	0.5239(6)	0.5605(2)
Avg. rank	3.89	2.75	3.67	5.29	4.12	**1.33**

Table 7
Subset Accuracy (↑) of six feature selection algorithms for DMLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.4892(3)	0.4815(4)	0.4940(2)	0.4766(6)	0.4769(5)	**0.4971(1)**
Flags	0.1395(4)	0.0774(6)	0.1503(2)	0.1042(5)	0.1450(3)	**0.1708(1)**
Emotions	0.2714(3)	0.2393(5)	0.2563(4)	0.1854(6)	0.2864(2)	**0.2947(1)**
CAL500	0.1938(4)	0.1941(3)	0.1908(6)	0.1925(5)	**0.1976(1)**	0.1956(2)
Yeast	0.1643(2)	0.1518(3.5)	0.1518(3.5)	0.1336(6)	0.1394(5)	**0.1655(1)**
Scene	0.5962(2)	0.4088(3)	0.5808(5)	0.4113(5)	0.5023(4)	**0.6261(1)**
Plant	**0.1124(1.5)**	0.1103(3)	0.0726(6)	0.0889(4)	0.0818(5)	**0.1124(1.5)**
WaterQuality	0.0160(4)	**0.0179(1)**	0.0169(3)	0.0075(6)	0.0132(5)	0.0170(2)
Gnegative	0.5349(4)	0.5204(5)	0.6440(2)	0.4466(6)	0.5923(3)	**0.6755(1)**
3sources	0.0142(2.5)	0.0142(2.5)	0.0113(6)	0.0141(4)	0.0114(5)	**0.0143(1)**
Human	0.0598(5)	0.0988(3)	0.0534(6)	0.0637(4)	0.1120(2)	**0.1146(1)**
Eukaryote	0.0462(4)	0.0489(3)	0.0551(2)	0.0267(6)	0.0277(5)	**0.0576(1)**
Avg. rank	3.25	3.5	4	5.37	3.76	**1.20**

rank list as the feature selection result. Consequently, the same number of features is selected with the quantity determined by *LDMI* as the final feature subset. In order to indicate the efficiency of *LDMI* compared with the other five algorithms, the *DMLkNN* and *MLkNN* classifier with six evaluation performances are used for experiments. Tables 2–13 display the classification performances of the two classifiers with six evaluation metrics on 12 data sets, which are obtained from the above proposed algorithms. The *bold* font in these tables are used to denote the best classification performance for each data set and the average rank for each algorithm, respectively. The number in parentheses represents the ranking of the algorithm on the current data set. The smaller the value, the higher the ranking, and the better the predictive classification performance.

As displayed from Tables 2–13, some observations can be obtained as follows:

1. For *DMLkNN*, in terms of *Hamming Loss* and *One Error*, the *LDMI* achieves superior performance for 11 data sets as compared to other algorithms, where the average performance of *LDMI* is less than the *RF-ML* in terms of *Hamming Loss* and *One Error*. It is known that *Hamming Loss* considers the label prediction with misclassified sample-label pairs, the proposed algorithm *LDMI* considers the label

Table 8
Hamming Loss (↓) of six feature selection algorithms for MLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.0511(6)	0.0487(3)	0.0470(1.5)	0.0505(5)	0.0496(4)	**0.0470(1.5)**
Flags	0.2706(2)	0.2823(4)	0.2829(5)	0.3100(6)	0.2787(3)	**0.2685(1)**
Emotions	0.2055(6)	0.2036(3)	0.2039(4)	0.2033(2)	0.2052(5)	**0.2027(1)**
CAL500	**0.1380(2)**	0.1387(4.5)	0.1389(6)	**0.1380(2)**	0.1387(4.5)	**0.1380(2)**
Yeast	0.2004(5)	0.2001(4)	0.2007(6)	0.1990(3)	0.1987(1.5)	**0.1987(1.5)**
Scene	0.0971(3)	0.0969(2)	0.1255(5)	0.1329(6)	0.1127(4)	**0.0941(1)**
Plant	0.0849(4.5)	**0.0827(1)**	0.0859(6)	0.0834(2)	0.0849(4.5)	0.0837(3)
WaterQuality	0.3043(5)	0.3030(4)	0.2998(2)	0.3251(6)	0.3027(3)	**0.2978(1)**
Gnegative	0.0702(4)	0.0732(5)	0.0588(2)	0.0794(6)	0.0652(3)	**0.0554(1)**
3sources	0.1884(6)	**0.1871(1.5)**	0.1880(4)	0.1880(4)	0.1880(4)	**0.1871(1.5)**
Human	0.6961(3.5)	0.7018(6)	0.6961(3.5)	0.6991(5)	0.6906(2)	**0.6904(1)**
Eukaryote	0.0513(4)	0.0511(3)	**0.0510(1.5)**	0.0516(6)	0.0515(5)	**0.0510(1.5)**
Avg. rank	4.25	3.42	3.87	4.42	3.62	**1.42**

Table 9
One Error (↓) of six feature selection algorithms for MLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.7329(5)	0.7235(4)	0.7175(3)	0.7470(6)	0.7174(2)	**0.7066(1)**
Flags	0.2113(6)	0.1966(5)	0.1792(2)	0.1955(4)	0.1900(3)	**0.1747(1)**
Emotions	0.2801(4.5)	0.2750(3)	0.2749(2)	0.2818(6)	0.2801(4.5)	**0.2733(1)**
CAL500	**0.1136(1.5)**	0.1176(5)	0.1176(5)	0.1176(5)	0.1175(3)	**0.1136(1.5)**
Yeast	0.2371(5)	0.2367(4)	0.2387(6)	0.2230(3)	0.2218(2)	**0.2214(1)**
Scene	0.2609(2)	0.2646(3)	0.3722(5)	0.3772(6)	0.3158(4)	**0.2489(1)**
Plant	0.6620(2.5)	0.6620(2.5)	0.6763(4)	0.6844(6)	0.6834(5)	**0.6446(1)**
WaterQuality	**0.2934(1)**	0.3047(4)	0.3025(3)	0.3472(6)	0.3189(5)	0.2978(2)
Gnegative	0.6547(4)	0.6397(3)	0.6160(2)	0.6677(6)	0.6612(5)	**0.6009(1)**
3sources	0.6877(3)	0.7018(6)	0.6906(4)	0.6991(5)	**0.6847(1)**	0.6876(2)
Human	0.6363(4)	0.6031(3)	0.6563(6)	0.6453(5)	0.6003(2)	**0.5996(1)**
Eukaryote	0.6235(2)	0.6374(4)	0.6242(3)	0.6566(6)	0.6516(5)	**0.6053(1)**
Avg. rank	3.38	3.25	3.8	5.33	3.54	1.21

Table 10
Coverage (↓) of six feature selection algorithms for MLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	2.8472(4)	2.7504(3)	2.8954(5)	2.9679(6)	2.3599(2)	**2.2850(1)**
Flags	**3.6689(1)**	3.8021(6)	3.7716(4)	3.7900(5)	3.6987(3)	3.6926(2)
Emotions	1.8402(2)	1.8518(4)	1.8438(3)	1.8840(6)	1.8574(5)	**1.8381(1)**
CAL500	**129.4387(1)**	130.2846(6)	129.5820(3)	129.9999(5)	129.8678(4)	129.5669(2)
Yeast	6.3984(5)	6.3682(4)	6.4343(6)	**6.2771(1)**	6.3652(3)	6.3557(2)
Scene	0.5305(2)	0.5567(3)	0.7312(5)	0.8052(6)	0.6951(4)	**0.5272(1)**
Plant	2.2004(6)	2.0093(2)	2.0462(3)	2.1862(5)	2.1749(4)	**2.0021(1)**
WaterQuality	9.0389(3)	9.0330(2)	9.0447(4)	9.4009(6)	9.2179(5)	**9.0321(1)**
Gnegative	0.5865(4)	0.5513(3)	0.5219(2)	0.6210(6)	0.5988(5)	**0.4811(1)**
3sources	**2.1290(1)**	2.1830(5)	2.1967(6)	2.1631(3)	2.1713(4)	2.1342(2)
Human	2.4101(5)	2.2814(3)	2.4735(6)	2.4056(4)	2.1941(2)	**2.1838(1)**
Eukaryote	3.0347(5)	2.8918(3)	2.7743(2)	3.0328(4)	3.0604(6)	**2.7368(1)**
Avg. rank	3.25	3.67	4.08	4.75	3.87	**1.33**

distribution with more supervised information and can effectively degrade the misclassification loss. Note that *LDMI* outperforms other algorithms on 10 data sets in terms of the *Coverage*, *Average Precision*, *Ranking Loss* and *Subset Accuracy*. It achieves a performance close to the best value on the other one data set. Specifically, in Table 4, *LDMI* achieves an average performance of 26.35% less than *RF-ML* in terms of *Coverage*, which indicates the superiority of label distributions. As for the average ranking, *LDMI* achieves the highest ranking than the lowest ranking, and the differences for six evaluation metrics between the best and the worst are 3.21, 3.83, 4.27, 3.89, 3.96, and 4.17, respectively. It is known that the significance among multi-labels may be different, which can influence the label ranking in terms of the *Average Precision* and *Ranking Loss*. The classification results demonstrate the superiority of considering the labeling-significance of each instance for multi-label classification.

2. For *MLkNN, LDMI* outperforms the other compared algorithms on most data sets. As displayed in Table 8, the *LDMI*, *MDDMspc* and *RF-ML* obtain the same predictive performance on *CAL500*, simultaneously. It is known that the class of label on *CAL500* reached 174 with the imbalanced data resulting in the worse performance, in close proximity. Particularly, the difference of the highest performance and the lowest performance reaches 3.9% in *One*

Table 11
Ranking Loss (↓) of six feature selection algorithms for MLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.1018(4)	0.0989(3)	0.1043(5)	0.1122(6)	0.0820(2)	**0.0783(1)**
Flags	0.2075(3)	0.2176(5)	**0.2031(1)**	0.2229(6)	0.2095(4)	0.2055(2)
Emotions	0.1719(3)	**0.1687(1)**	0.1720(4)	0.1766(6)	0.1764(5)	0.1718(2)
CAL500	0.1834(3)	0.1832(2)	0.1836(4)	0.1845(6)	0.1837(5)	**0.1830(1)**
Yeast	0.1750(5)	0.1735(4)	0.1776(6)	0.1715(2)	0.1720(3)	**0.1714(1)**
Scene	0.0887(2)	0.0937(3)	0.1290(5)	0.1437(6)	0.1221(4)	**0.0882(1)**
Plant	0.1884(6)	0.1737(2)	0.1755(3)	0.1860(5)	0.1810(4)	**0.1729(1)**
WaterQuality	0.2725(3)	0.2719(2)	0.2819(4)	0.3049(6)	0.2849(5)	**0.2697(1)**
Gnegative	0.0773(4)	0.0719(3)	0.0686(2)	0.0824(6)	0.0793(5)	**0.0630(1)**
3sources	0.4007(2)	0.4069(5)	0.4049(4)	0.4071(6)	0.4022(3)	**0.4005(1)**
Human	0.1638(5)	0.1508(2)	0.1688(6)	0.1637(4)	0.1509(3)	**0.1463(1)**
Eukaryote	0.1309(5)	0.1259(3)	0.1230(2)	0.1306(4)	0.1320(6)	**0.1182(1)**
Avg. rank	3.75	2.91	3.83	5.25	4.08	**1.17**

Table 12
Average Precision (↑) of Five feature selection algorithms for MLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.5667(4)	0.5792(3)	0.5631(5)	0.5239(6)	0.5998(2)	**0.6212(1)**
Flags	0.8131(3)	**0.8185(1)**	0.8124(4)	0.7991(6)	0.8042(5)	0.8151(2)
Emotions	0.7910(4)	0.7913(3)	0.7926(2)	0.7877(6)	0.7901(5)	**0.7936(1)**
CAL500	0.4913(5)	0.4916(3)	0.4916(3)	0.4890(6)	0.4916(3)	**0.4919(1)**
Yeast	0.7547(5)	0.7560(4)	0.7501(6)	**0.7586(2.5)**	0.7591(1)	**0.7586(2.5)**
Scene	0.8454(2)	0.8401(3)	0.7778(5)	0.7679(6)	0.8059(4)	**0.8511(1)**
Plant	0.5572(3)	0.5589(2)	0.5424(6)	0.5511(4)	0.5502(5)	**0.5613(1)**
WaterQuality	0.6761(4)	0.6768(3)	0.6677(5)	0.6392(6)	0.6817(2)	**0.6847(1)**
Gnegative	0.7228(4)	0.7393(3)	0.7623(2)	0.7094(6)	0.7191(5)	**0.7787(1)**
3sources	0.5072(4)	0.5041(5)	0.5103(3)	0.5009(6)	**0.5127(1)**	0.5112(2)
Human	0.5897(2)	0.5799(3)	0.5562(5)	0.5529(6)	0.5619(4)	**0.5926(1)**
Eukaryote	0.5315(6)	0.5457(5)	0.5570(3)	0.5602(2)	0.5468(4)	**0.5737(1)**
Avg. rank	3.87	3.17	4.08	5.13	3.41	**1.29**

Table 13
Subset Accuracy (↑) of six feature selection algorithms for MLkNN.

Data sets	MDDMspc	MDDMproj	PMU	RF-ML	MLFRS	LDMI
Birds	0.4861(5)	0.4956(2)	0.4923(3)	0.4769(6)	0.4877(4)	**0.4971(1)**
Flags	0.1342(2)	0.0782(6)	0.1289(4)	0.1042(5)	0.1339(3)	**0.1437(1)**
Emotions	0.2613(3)	0.2342(5)	0.2359(4)	0.1871(6)	0.2746(2)	**0.2829(1)**
CAL500	0.1945(4)	0.1954(3)	0.1924(6)	0.1942(5)	0.1990(1.5)	**0.1990(1.5)**
Yeast	0.1419(5)	0.1514(3)	0.1510(4)	0.1407(6)	0.1618(2)	**0.1630(1)**
Scene	0.5858(2)	0.5791(3)	0.4101(5)	0.4084(6)	0.5023(4)	**0.6207(1)**
Plant	0.0726(6)	**0.1206(1)**	0.0879(4)	0.1032(3)	0.0818(5)	0.1124(2)
WaterQuality	0.0160(3.5)	0.0170(1.5)	0.0160(3.5)	0.0085(6)	0.0132(5)	**0.0170(1.5)**
Gnegative	0.5349(4)	0.5161(5)	0.6411(2)	0.4466(6)	0.5916(3)	**0.6719(1)**
3sources	0.0114(2.5)	0.0114(2.5)	0.0113(5)	0.0113(5)	0.0113(5)	**0.0142(1)**
Human	**0.1027(1)**	0.0969(3)	0.0541(5.5)	0.0634(4)	0.0541(5.5)	0.1005(2)
Eukaryote	0.0469(4)	**0.0586(1)**	0.0492(3)	0.0267(5)	0.0257(6)	0.0552(2)
Avg. rank	3.50	3.09	4.08	5.25	3.83	**1.33**

Error where the best performance and second best differed by 1.57%. Subsequently, it is because *LDMI* proposed in this study considering different labeling-significance can effectively affect the top-ranked label, which is not the vector of proper labels in terms of *One Error*. To sum up, the difference is approximately 18.1%. Similarly, both in *DMLkNN* and *MLkNN*, the proposed algorithm *LDMI* outperforms other state-of-the-art multi-label feature selection algorithms, which indicates the effectiveness when considering the label significance degree for label distribution feature selection.

For a reliable and dependable validation of the performance comparison among the compared algorithms, we conducted extensive experiments to display the change tendency with different number of feature subset for each classification performance explored. The increasing step size of each feature subset for *Birds* is 3 and *Scene* is 4. Figs. 5–8 display the different evaluation classification performance on *Birds* and *Scene*. In these figures, the horizontal and vertical represent the size of the selected features and the classification performance. There are six lines in each figure, corresponding to *LDMI*, *MDDMspc*, *MDDMproj*, *PMU*, *RF-ML* and *MLFRS*, respectively. As shown in Figs. 5–8, it can be observed that each figure of evaluation metrics does not decrease or increase monotonously with the number of selected features.

Fig. 5. Six evaluation metrics by the number of selected features on Birds from DMLkNN.

LDMI can obtain optimal classification performance as the number of features selected increase, no matter how the variation tendency of the curves. From Figs. 5 and 6, *LDMI* outperforms the other compared algorithms in most cases. Especially, in terms of *Hamming Loss* and *One Error*, *LDMI* achieves superior performance when the number of selected features is between 9 and 60. In addition, for *Scene*, though the classification performance of *LDMI* is close to that of *MDDMspc* and *MDDMproj*, most of metrics have optimal values especially for *Ranking Loss* when the number of selected features is between 20 and 60. Note that *LDMI* will get better classification performance with a certain number of features, and it meets the actual situation. The better performance of *LDMI* indicates the effectiveness of utilizing the significance between each label.

Fig. 6. Six evaluation metrics by the number of selected features on Birds from MLkNN.

5.4.2. Evaluation of different label enhancement algorithms

In this subsection, a comparison of different label enhancement algorithms is conducted. Two of representative label enhancement algorithms, e.g., *GLLE* [68], *LIABLE* [19], are selected for comparing with our proposed label enhancement *LEMN*. Parameter settings of each method are recommended in [19,68]. To compare the effectiveness of each algorithm, six multi-label evaluation metrics on six representative data sets are compared and analyzed using feature selection algorithm (*LDMI*) proposed in this paper after label enhancement. Additionally, to better show the experimental results, Coverage on data set *CAL500* is scaled proportionally. The results of six evaluation metrics from two classifiers are shown in Figs. 9 and 10 where ↓ denote the

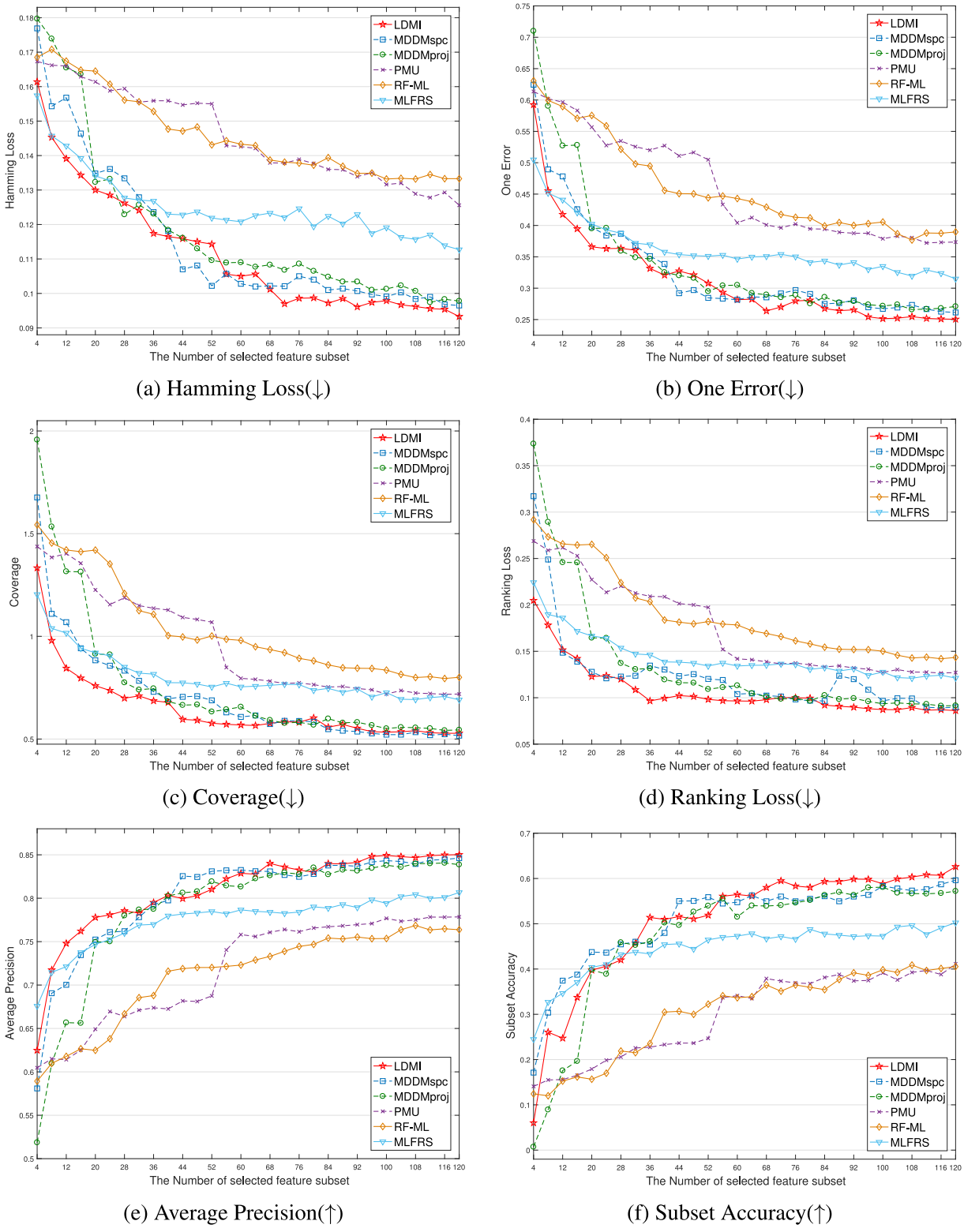


Fig. 7. Six evaluation metrics by the number of selected features on Scene from DMLkNN.

smaller the value, the better the performance, and ↑ denote the larger the value, the better the performance.

For DMLkNN in Fig. 9, *LEMN* outperforms the other algorithms on 4 data sets of *Hamming Loss*, *Coverage*, *Ranking Loss* and *Average Precision*. Particularly on data set *Birds*, *LEMN* obtain a superior performance on all evaluation metrics. It is determined that the label enhancement algorithm designed in this paper can effectively transform the logical label into label distribution. Generating the neighborhood relationship among data, and jointly local relevance and global relevance is benefit for label enhancement. Additionally, the features selected algorithm *LDMI* can be applying to achieve better classification performance.

In terms of the MLkNN in Fig. 10, *LEMN* obtains the best performance on 66.67% data sets of *One Error*, *Coverage*, *Ranking*

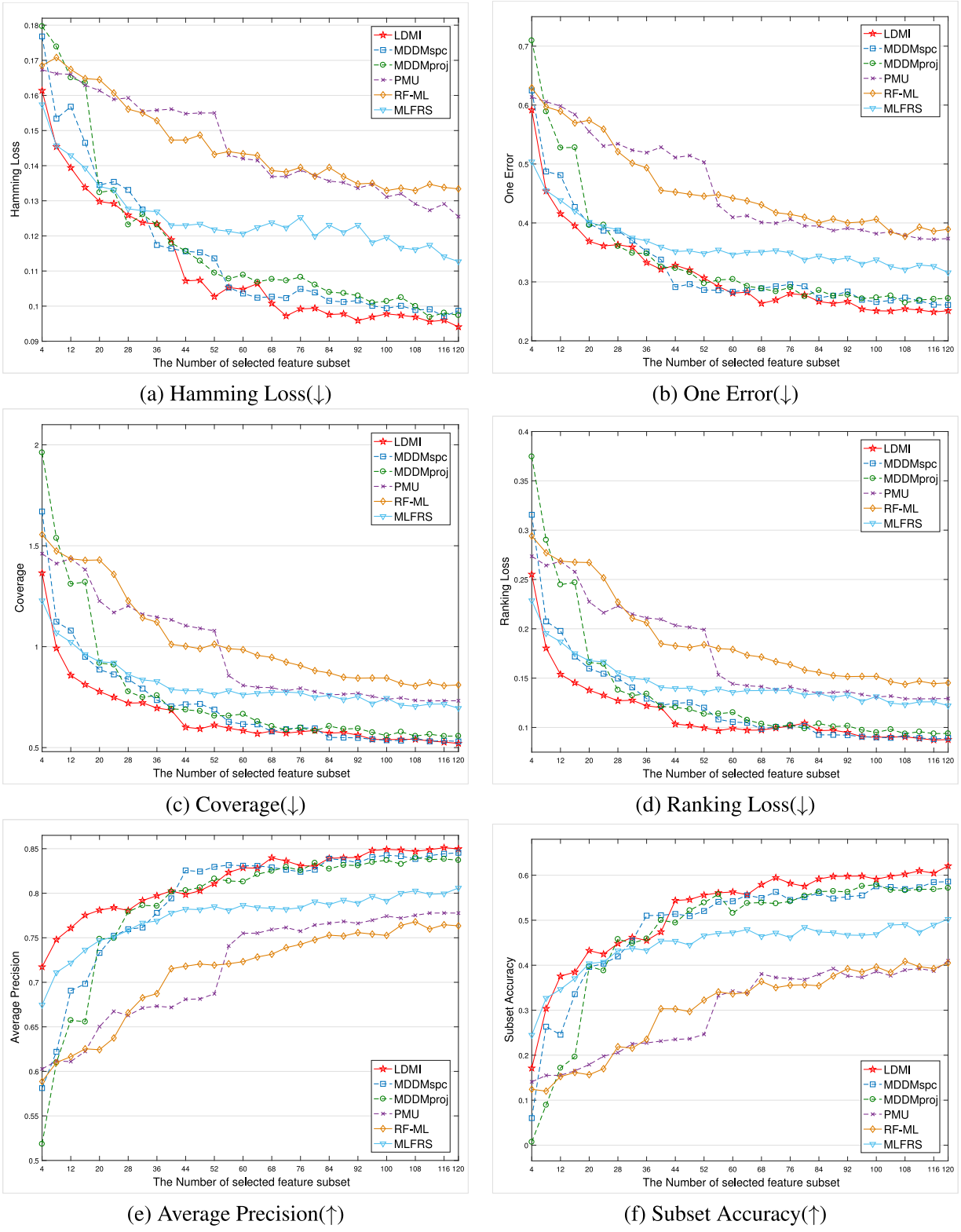


Fig. 8. Six evaluation metrics by the number of selected features on Scene from MLkNN.

Loss, *Average Precision* and *Subset Accuracy*. For all six data sets, *LEMN* obtain the superior performance of six evaluation metrics on data sets *Birds*, *Emotions* and *Flags* in most cases. It can be determined that *LEMN* embedded in label distribution feature selection algorithm *LDMI* can achieve better effectiveness on middle or high-dimensional data. In particular, three algorithms obtain a extremely closed performance for all six evaluation metrics on data set *CAL500*. As it is known that the labels of *CAL500* is 174, which demonstrates that too many labels can result in a badly effectiveness of label enhancement.

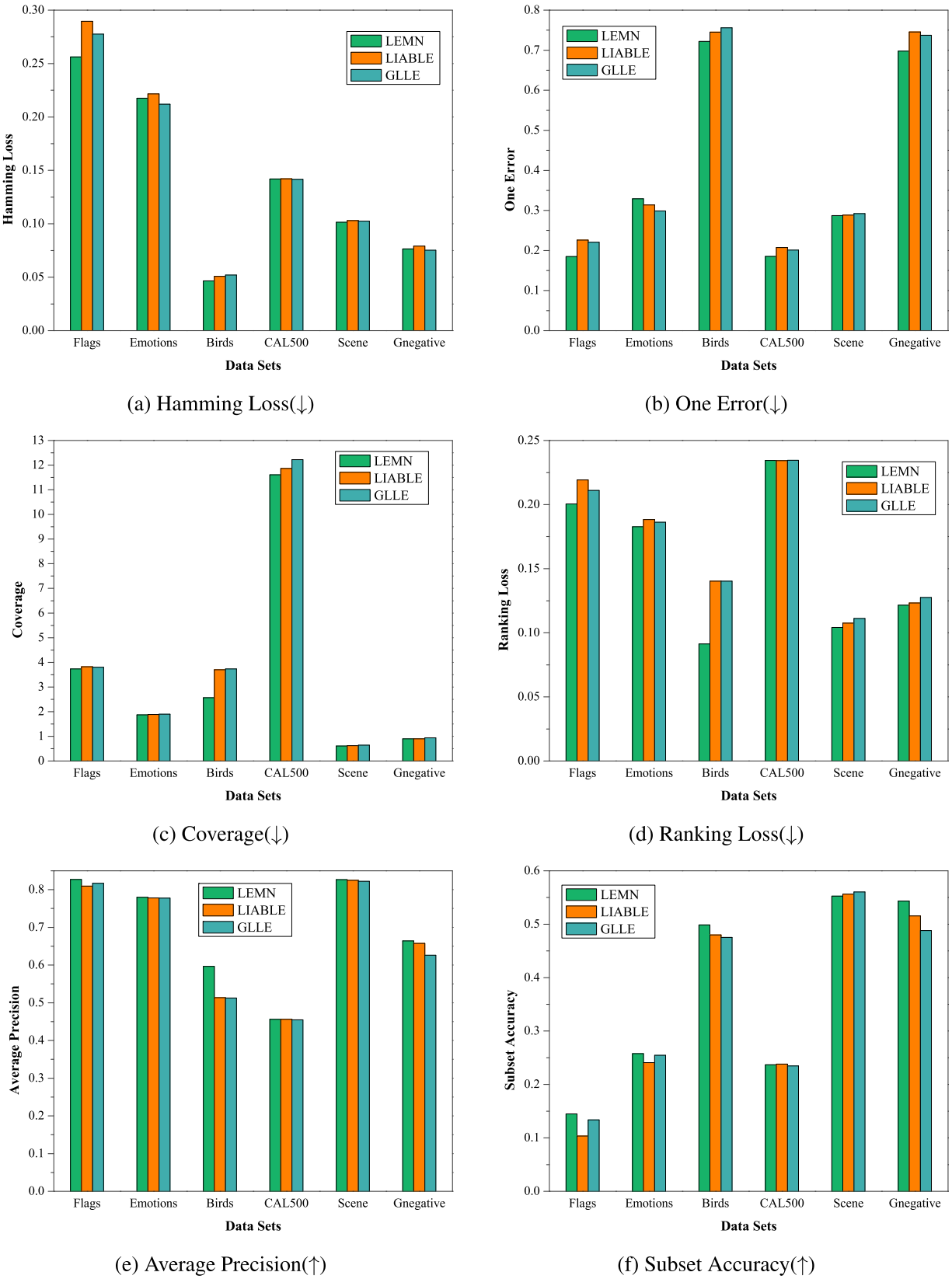


Fig. 9. Six evaluation metrics of three label enhancement algorithms on six data sets from DMLkNN.

As the aforementioned analysis, the classification performances of *LEMN* represented based on random variable distribution is better than the other two label enhancement algorithms in some cases, since the feature selection algorithm *LDMI* is proposed based on the LEMN algorithm. In summary, the proposed label distribution feature selection algorithm *LDMI* contributes

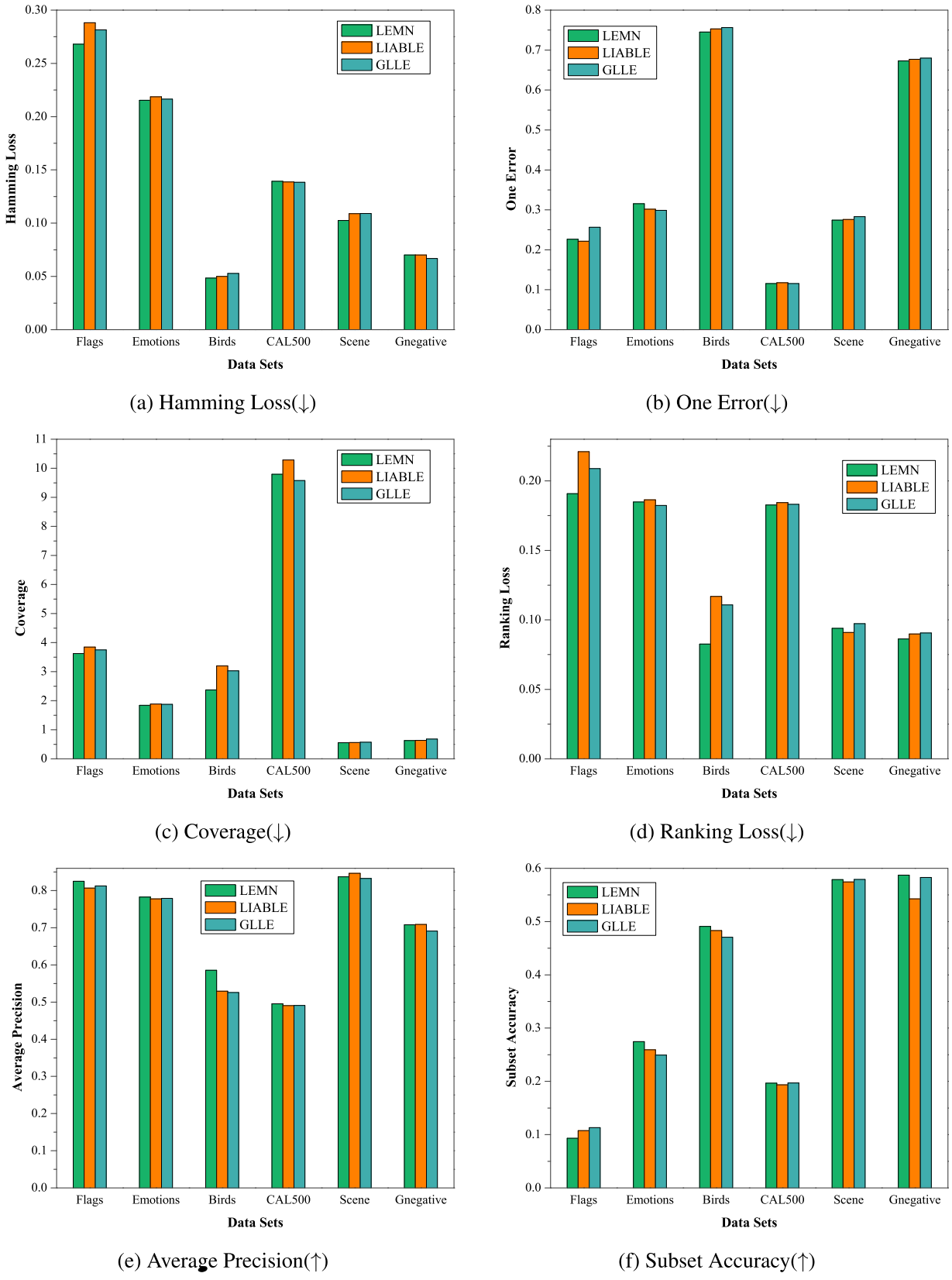


Fig. 10. Six evaluation metrics of three label enhancement algorithms on six data sets from MLkNN.

to consider and optimize the different significance of each related label for traditional multi-label learning based on label enhancement, and effectively employing this more supervised information to improve the classification performance.

5.4.3. Evaluation of the statistical test

Friedman statistical test: The Friedman test [69] is used for comparing each classification performance analysis to those of other compared algorithms systematically. It is a popular statistical test

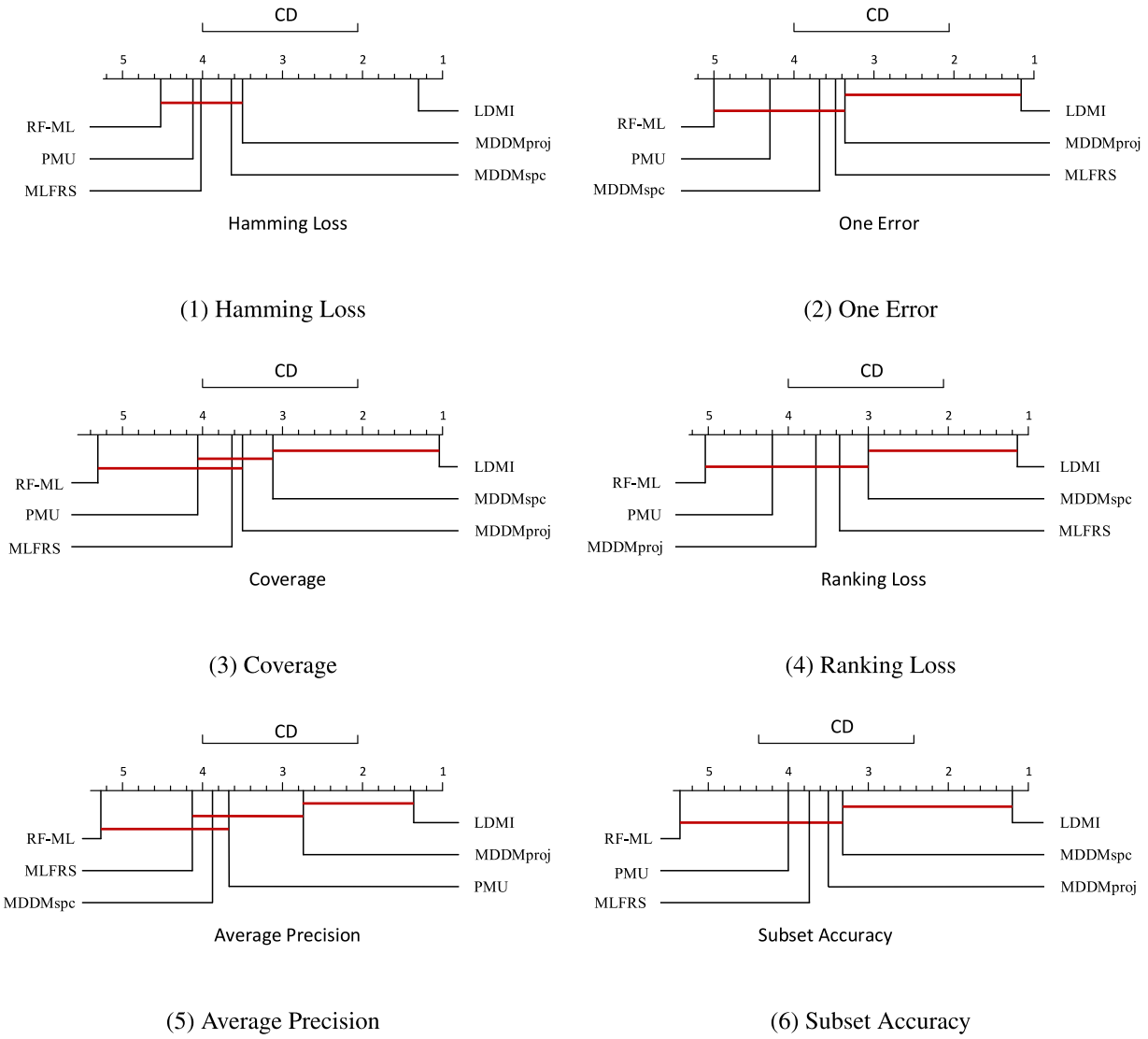


Fig. 11. Nemenyi test results of all compared methods for six evaluation metrics from DMLkNN.

for machine learning with a wide application. Assuming that k and N are given, where k is the number of compared algorithms and N is the number of data sets. Given the average ranks of algorithms $R_j = \frac{1}{N}\sum_i r_i^j$ among all data sets:

$$\chi_F^2 = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}]$$

according to the F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

The F_F for different classification performances for two classifiers are displayed in Table 14, which shows that the compared algorithms are rejected under the *Friedman test* where the significance $\alpha = 0.05$. The compared algorithms exhibited remarkably equivalent performance.

Nemenyi test: A post-hoc test namely *Nemenyi test* is used to analyze the relative performance for all the compared algorithms. In addition, the formula of critical difference is defined as follow:

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}}$$

Table 14
Different evaluation metrics of the Friedman statistics F_F ($k = 6$, $N = 12$) and the Critical value($\alpha = 0.05$).

Evaluation metrics	$F_F(DMLkNN)$	$F_F(MLkNN)$	Critical value
Hamming Loss	6.7772	5.5848	
One Error	7.9623	4.8195	
Coverage	10.4105	6.4779	2.3828
Ranking Loss	3.7684	12.4108	
Average Precision	12.7388	8.8072	
Subset Accuracy	13.7652	11.5009	

Given $k = 6$, $N = 12$, $\alpha = 0.05$, we can obtain $q_\alpha = 2.589$ and $CD = 2.17$. The results on different evaluation metrics are displayed in Figs. 11 and 12, where the average rank of each compared algorithm is plotted along the axis. In the sub-figures, it will be inter-connected with a thick line, which is the average ranks between the proposed and other compared algorithms. This implies that the proposed method achieves a better performance as compared to other algorithms.

Based on the results shown in Figs. 11 and 12, some observations can be obtained as follow. The proposed algorithm *LDMI* significantly outperforms other compared algorithms. For

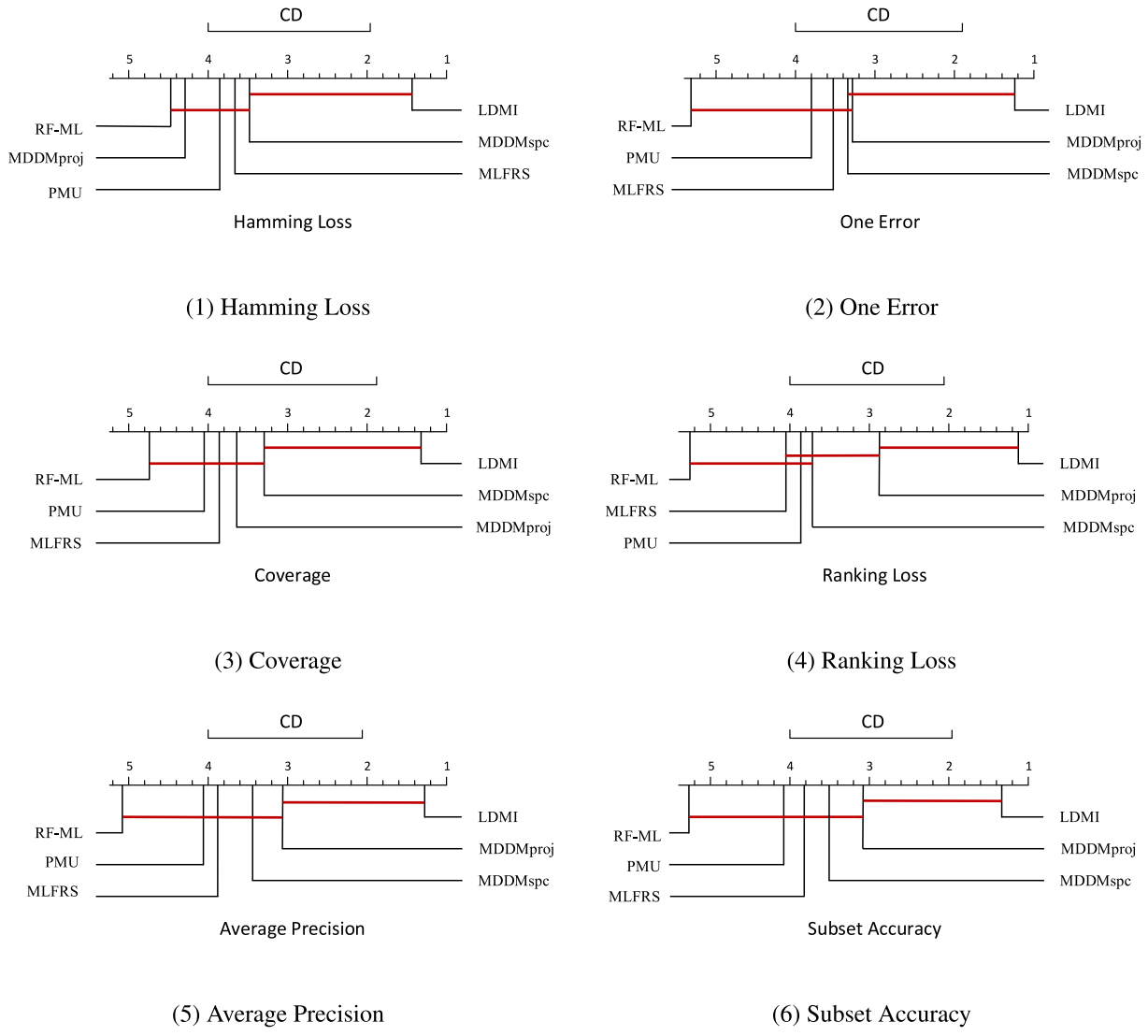


Fig. 12. Nemenyi test results of all compared methods for six evaluation metrics from MLkNN.

DMLkNN, LDMI achieves a comparable performance against the other compared algorithms in terms of *Hamming Loss*. With respect to *One Error* and *Average Precision*, there is no consistent evidence to indicate statistical differences between *LDMI* and *MDDMproj*, and there is no consistent evidence to indicate statistical differences between *LDMI* and *MDDMspc* in terms of *Coverage, Ranking Loss* and *Subset Accuracy*. *LDMI* achieves statistically superior performance against *RF-ML*, *PMU* and *MLFRS*. For *MLkNN*, *LDMI* outperforms most of the other compared algorithms statistically using the six evaluation metrics, which further demonstrate that the feasibility of label significance for multi-label classification. In terms of *Hamming Loss* and *Coverage*, there is no consistent evidence to indicate the statistical differences between *LDMI* and *MDDMspc* where *LDMI* outperforms the other compared algorithms. In terms of other four evaluation metrics, *LDMI* achieves statistically superior performance compared with *RF-ML*, *PMU*, *MLFRS* and *MDDMspc*. To summarized, *LDMI* obtains superior competitive performance when compared with the state-of-art multi-label feature selection algorithms, which indicates the effectiveness of label significance for label distribution feature selection.

6. Conclusion and future work

Feature selection is necessary to reduce the high-dimensionality in multi-label learning. Although some algorithms have already been proposed, most existing multi-label feature selection methods do not consider the significance of related label for each instance. Therefore, we first make use of the related labels based on label distribution learning to select a feature subset for the multi-label data. A measurement using random variable distribution is designed to convert the logical labels in the multi-label data into label distribution with more supervised information. Subsequently, a label distribution feature selection based on mutual information is proposed, which can remove some redundant or irrelevant features. Finally, the experimental results demonstrate that the proposed algorithms are more efficient than some state-the-art methods.

Although this study provides a solution to multi-label feature selection problem, there are some open directions or research issues. Based on the aforementioned results, some further investigations are as follows:

1. In our current study, we considers the inter-significance of different related labels for every instance in terms of continuous numerical data, more complex data or more

practical applications for multi-label learning will be considered. Thereby we will focus on the multi-source data [70], multi-view data [3] and the fuzzy rough set [5]. In terms of real applications, we will focus on the image or video annotation [49], and the electrical or the economic problem [71].

2. Neighborhood granularity is applied to our current works, while the initialization of the some proposed optimization scheme from clustering techniques is typically a good start for iterative optimization, which can be replaceable to improve the rate of convergence [4,31].
3. Some optimization algorithms embedded in feature selection can be used to achieve superior feature subset, such as population extremal optimization [72], multi-objective extremal optimization [73], and the critical infrastructures combined with granular computing [57].
4. Besides, in our works, although we applied the labeling-significance to multi-label learning to employ more supervised information, more general case in label distribution learning will be concerned in future works.

CRediT authorship contribution statement

Wenbin Qian: Conceptualization, Methodology, Writing - original draft, Visualization, Funding acquisition. **Jintao Huang:** Software, Data curation, Writing - review & editing. **Yinglong Wang:** Supervision, Visualization. **Wenhao Shu:** Validation, Formal analysis.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No. 61966016 and 61662023), the Natural Science Foundation of Jiangxi Province, China (No. 20192BAB207018), the Scientific Research Project of Education department of Jiangxi Province, China (No. GJJ180200), and the Graduate Innovation Special Fund Project of Jiangxi Province, China (No. YC2018-S192).

References

[1] Y.Y. Yao, Three-way decision and granular computing, Internat. J. Approx. Reason. 103 (2018) 107–123.
[2] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, Positive approximation: An accelerator for attribution reduction in rough set theory, Artificial Intelligence 174 (2010) 597–618.
[3] H. Wang, Y. Yang, B. Liu, H. Fujita, A study of graph-based system for multi-view clustering, Knowl.-Based Syst. 163 (2019) 1009–1019.
[4] Y.L. Zhang, Y. Yang, T.R. Li, H. Fujita, A multitask multiview clustering algorithm in heterogeneous situations based on LLE and LE, Knowl.-Based Syst. 163 (2019) 776–786.
[5] H. Fujita, A. Gaeta, V. Loia, F. Orciuoli, Hypotheses analysis and assessment in counter-terrorism activities: a method based on OWA and Fuzzy Probabilistic Rough Sets, IEEE Trans. Fuzzy Syst. (2019) http://dx.doi.org/10.1109/TFUZZ.2019.2955047.
[6] H.M. Chen, T.R. Li, Y. Cai, C. Luo, H. Fujita, Parallel attribute reduction in dominance-based neighborhood rough set, Inform. Sci. 373 (2016) 351–368.
[7] Z.H. Jiang, K.Y. Liu, X.B. Yang, H.L. Yu, H. Fujita, Y.H. Qian, Accelerator for supervised neighborhood based attribute reduction, Internat. J. Approx. Reason. 119 (2020) 122–150.
[8] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (2014) 1819–1837.
[9] S. Kashef, H. Nezamabadi-pour, B. Nikpour, Multi-label feature selection: A comprehensive review and guiding experiments, Wires Data Min. Knowl. Discov. 8 (2018) 1–29.
[10] Z.H. Zhou, M.L. Zhang, S.J. Huang, Y.F. Li, Multi-instance multi-label learning, Artificial Intelligence 176 (2012) 2291–2320.
[11] N.T. Rubin, A. Chambers, P. Smyth, Statistical topic models for multi-label document classification, Mach. Learn. 88 (2012) 157–208.
[12] A.B. Samle, A. Masri, Shahrul A.M. Noah, Feature ranking for enhancing boosting-based multi-label text categorization, Expert Syst. Appl. 113 (2018) 531–543.
[13] Y. Liu, K.W. Wen, Q.X. Gao, et al., SVM based multi-label learning with missing labels for image annotation, Pattern Recognit. 78 (2018) 307–317.
[14] W.J. Chen, Y.H. Shao, C.N. Li, N.Y. Deng, MLTSVM: A novel twin support vector machine to multi-label learning, Pattern Recognit. 52 (2016) 61–74.
[15] Y. Zhu, K.M. Ting, Z.H. Zhou, Multi-label learning with emerging new labels, IEEE Trans. Knowl. Data Eng. 30 (2018) 1901–1914.
[16] Z.F. He, M. Yang, Sparse and low-rank representation for multi-label classification, Appl. Intell. 49 (5) (2019) 1708–1723.
[17] X. Geng, Label distribution learning, IEEE Trans. Knowl. Data Eng. 28 (2016) 1731–1748.
[18] C.D. Xu, X. Geng, Hierarchical classification based on label distribution learning, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI'19, 2019, pp. 5533–5540.
[19] Y.K. Li, M.L. Zhang, X. Geng, Leveraging implicit relative labeling-importance information for effective multi-label learning in: Proceedings of IEEE Intelligent Conference on Data Mining, 2015, pp. 251–260.
[20] X.D. Yue, Y.F. Chen, D.Q. Miao, H. Fujita, Fuzzy neighborhood covering for three-way classification, Inform. Sci. 507 (2020) 795–808.
[21] Y.J. Lin, Y.W. Li, C.X. Wang, Attribute reduction for multi-label learning with fuzzy rough set, Knowl. Based Syst. 152 (2017) 51–61.
[22] X.B. Yang, Y.Y. Yao, Ensemble selector for attribute reduction, Appl. Soft Comput. 70 (2018) 1–11.
[23] W. Weng, Y.J. Lin, S.X. Wu, et al., Multi-label learning based on label-specific features and local pairwise label correlation, Neurocomputing 273 (2018) 385–394.
[24] Z.F. He, M. Yang, Y. Gao, et al., Joint multi-label classification and label correlations with missing labels and feature selection, Knowl.-Based Syst. 163 (2019) 145–158.
[25] J.H. Xu, A weighted linear discriminant analysis framework for multi-label feature extraction, Neurocomputing 275 (2018) 107–120.
[26] P.F. Zhu, Q. Xu, Q.H. Hu, et al., Multi-label feature selection with missing labels, Pattern Recognit. 74 (2018) 488–502.
[27] X.Y. Zhang, D.Q. Miao, Quantitative/qualitative region-change uncertainty/certainty in attribute reduction: Comparative region-change analyses based on granular computing, Inform. Sci. 334 (335) (2016) 174–204.
[28] X.Y. Jia, L. Shang, B. Zhou, Y.Y. Yao, Generalized attribute reduction in rough set theory, Knowl.-Based Syst. 91 (2016) 204–218.
[29] K. Mikalsen, C.S. Ruizb, F.M. Bianchi, R. Jenssen, Noisy multi-label semi-supervised dimensionality reduction, Pattern Recognit. 90 (2019) 257–270.
[30] Y.J. Lin, Q.H. Hu, J.H. Liu, Streaming feature selection for multi label learning based on fuzzy mutual information, IEEE Trans. Fuzzy Syst. 25 (2017) 1491–1507.
[31] F. Li, D.Q. Miao, W. Pedrycz, Granular multi-label feature selection based on mutual information, Pattern Recognit. 67 (2017) 410–423.
[32] H.B. Yu, T. Zhang, W.J. Wen, Shared subspace least squares multi-label linear discriminant analysis, Appl. Intell. 49 (1) (2019) 1–12.
[33] H.K. Lim, J.S. Lee, D.W. Kim, Optimization approach for feature selection in multi-label classification, Pattern Recognit. Lett. 89 (2017) 25–30.
[34] S.P. Xu, X.B. Yang, H.L. Yu, D.J. Yu, Multi-label learning with label-specific feature reduction, Knowl.-Based Syst. 104 (2016) 52–61.
[35] Y.W. Wang, L.Z. Feng, A new hybrid feature selection based on multi-filter weights and multi-feature weights, Appl. Intell. 49 (12) (2019) 4033–4057.
[36] M. Juan, G. R.Daniela, Z. Alejandro, An empirical comparison of feature selection methods in problem transformation multi-label classification, IEEE Lat. Am. Trans. 14 (2016) 3784–3791.
[37] X.D. Wang, R.C. Chen, C.Q. Hong, Z.Q. Zeng, Z.L. Zhou, Semi-supervised multi-label feature selection via label correlation analysis with l1-norm graph embedding, Image Vis. Comput. 63 (2017) 10–23.
[38] C.X. Wang, Y.J. Lin, J.H. Liu, Feature selection for multi-label learning with missing labels, Appl. Intell. 49 (8) (2019) 3027–3042.
[39] J.H. Ma, T.W.S. Chow, Robust non-negative sparse graph for semi-supervised multi-label learning with missing labels, Inform. Sci. 422 (2018) 336–351.
[40] S.B. Chen, Y.M. Zhang, C.H.Q. Ding, et al., Extended adaptive lasso for multi-class and multi-label feature selection, Knowl.-Based Syst. 173 (2019) 28–36.
[41] J.H. Liu, Y.J. Lin, S.X. Wu, C.X. Wang, Online multi-label group feature selection, Knowl.-Based Syst. 143 (2018) 42–57.
[42] J.H. Liu, Y.J. Lin, Y.W. Li, et al., Online multi-label streaming feature selection based on neighborhood rough set, Pattern Recognit. 84 (2018) 273–287.
[43] S. Kashef, H. Nezamabadi. P., A label-specific multi-label feature selection algorithm based on the Pareto dominance concept, Pattern Recognit. 88 (2019) 654–667.

[44] W. Weng, Y.J. Lin, S.X. Wu, et al., Multi-label learning based on label-specific features and local pairwise label correlation, Neurocomputing 273 (2018) 385–394.
[45] N. Spolaor. E. Cherman, M. Monard, ReliefF for multi-label feature selection, in: Brazilian Conference on Intelligent Systems, 2013, pp. 6–11.
[46] J. Zhang, C.D. Li, D.L. Cao, et al., Multi-label learning with label-specific features by resolving label correlations, Knowl.-Based Syst. 159 (2018) 148–157.
[47] X. Geng, C. Yin, Z.H. Zhou, Facial age estimation by label distribution learning, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2018) 2401–2412.
[48] M.T. Chen, X.G. Wang, B. Feng, W.Y. Liu, Structured random forest for label distribution learning, Neurocomputing 320 (2018) 171–182.
[49] Z.X. Zhang, M. Wang, X. Geng, Crowd counting in public video surveillance by label distribution learning, Neurocomputing 166 (2015) 151–163.
[50] P. Zhao, Z.H. Zhou, Label distribution learning by optimal transport, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI'18, 2018, pp. 4506–4513.
[51] X. Geng, K. Smith-Miles, Z.H. Zhou, Facial age estimation by learning from label distribution, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence, AAAI'10, 2010, pp. 451–456.
[52] M. Xu, Z.H. Zhou, Incomplete label distribution learning, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17, 2017, pp. 3175–3181.
[53] B.B. Gao, C. Xing, C.W. Xie, J.X. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Trans. Image Process. 26 (2017) 2825–2838.
[54] J.F. Yang, L.Y. Chen, L. Zhang, et al., Historical context-based style classification of painting images via label distribution learning, in: Proceedings of the 26th ACM international conference on Multimedia, MM '18, 2018, pp. 1154–1162.
[55] T. Liu, A. Venkatachalam, P.S. Bongale, C. Homan, Learning to predict population-level label distributions, in: Companion Proceedings of The 2019 World Wide Web Conference Pages, WWW '19, 2019, pp. 1111–1120.
[56] J.D. Li, K.W. Cheng, S.H. Wang, et al., Feature selection: A data perspective, ACM Comput. Surv. 9 (39) (2010) 1–44.
[57] H. Fujita, A. Gaeta, V. Loia, F. Orciuoli, Resilience analysis of critical infrastructures: A cognitive approach based on granular computing, IEEE Trans. Cybern. 49 (5) (2019) 1835–1848.
[58] Y.H. Qian, J.Y. Liang, C.Y. Dang, Incomplete multigranulation rough set, IEEE Trans. Syst. 40 (2010) 420–431.
[59] Q. Wang, Y.H. Qian, X.Y. Liang, et al., Local neighborhood rough set, Knowl.-Based Syst. 153 (2018) 53–64.
[60] J.H. Li, Y. Ren, C.L. Mei, Y.H. Qian, X.B. Yang, A comparative study of multi-granulation rough sets and concept lattices via rule acquisition, Knowl.-Based Syst. 91 (2016) 152–164.
[61] Q.H. Hu, D. Y, Z.X. Xie, Neighborhood classifiers, Expert Syst. Appl. 34 (2008) 866–876.
[62] J. Wang, X. Geng, Theoretical Analysis of label distribution learning, in: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI'19, 2019, pp. 5256–5263.
[63] N. Spola, E. Cherman, M. Monard, H. Lee, Filter approach feature selection methods to support multi-label learning based on ReliefF and Information Gain, in: Proceedings of Brazilian Symposium on Artificial Intelligence, 2012, pp. 1–10.
[64] J. Read, B. Pfahringer, G. Holmes, et al., Classifier chains for multi-label classification, Mach. Learn. 85 (2011) 333–359.
[65] J. Lee, D.W. Kim, Feature selection for multi-label classification using multi-variate mutual information, Pattern Recognit. Lett. 34 (2013) 349–357.
[66] Y. Zhang, Z. Zhou, Multi-label dimensionality reduction via dependence maximization, ACM Trans. Knowl. Discov. Database 4 (2010) 1–21.
[67] O. Reyes, C. Morell, S. Ventura, Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context, Neurocomputing 161 (2015) 168–182.
[68] R.F. Shao, N. Xu, X. Geng, Multi-label learning with label enhancement, in: Proceedings of the 2018 IEEE International Conference on Data Mining, ICDM'18, 2018, pp. 437–446.
[69] J. Demsar, Statistical comparisons of classifiers over multiple data sets, Mach. Learn. 7 (2006) 1–30.
[70] B.B. Sang, Y.T. Guo, D.R. Shi, W.H. Xu, Decision-theoretic rough set model of multi-source decision systems, Int. J. Mach. Learn. Cybern. 9 (2018) 1941–1954.
[71] M.R. Chen, G.Q. Zeng, K.D. Lu, Constrained multi-objective population extremal optimization based economic-emission dispatch incorporating renewable energy resources, Renew. Energy 143 (2019) 277–294.
[72] K.D. Lu, W.N. Zhou, G.Q. Zeng, Y.Y. Zheng, Constrained population extremal optimization-based robust load frequency control of multi-area interconnected power system, Int. J. Electr. Power Energy Syst. 105 (2019) 249–271.
[73] G.Q. Zeng, J. Chen, Y.X. Dai, L.M. Li, C.W. Zheng, M.R. Chen, Design of fractional order PID controller for automatic regulator voltage system based on multi-objective extremal optimization, Neurocomputing 160 (2015) 173–184.