Streaming Feature Selection for Multilabel Learning Based on Fuzzy Mutual Information

Yaojin Lin, Qinghua Hu, *Senior Member, IEEE*, Jinghua Liu, Jinjin Li, and Xindong Wu, *Fellow, IEEE*

***Abstract*—Due to complex semantics, a sample may be associated with multiple labels in various classification and recognition tasks. Multilabel learning generates training models to map feature vectors to multiple labels. There are several significant challenges in multilabel learning. Samples in multilabel learning are usually described with high-dimensional features and some features may be sequentially extracted. Thus, we do not know the full feature set at the beginning of learning, referred to as streaming features. In this paper, we introduce fuzzy mutual information to evaluate the quality of features in multilabel learning, and design efficient algorithms to conduct multilabel feature selection when the feature space is completely known or partially known in advance. These algorithms are called multilabel feature selection with label correlation (MUCO) and multilabel streaming feature selection (MSFS), respectively. MSFS consists of two key steps: online relevance analysis and online redundancy analysis. In addition, we design a metric to measure the correlation between the label sets, and both MUCO and MSFS take label correlation to consideration. The proposed algorithms are not only able to select features from streaming features, but also able to select features for ordinal multilabel learning. However streaming feature selection is more efficient. The proposed algorithms are tested with a collection of multilabel learning tasks. The experimental results illustrate the effectiveness of the proposed algorithms.**

***Index Terms*—Feature selection, fuzzy mutual information, label correlation, multilabel learning, streaming features.**

I. Introduction

In CLASSICAL supervised learning, each object only belongs to one of the candidate classes. This is inapplicable to some real-world applications [3], [11], [16], [17], [40], [43], [57]. In fact, some objects are associated with multiple concepts simultaneously. For example, a newspaper article concerning the reactions of the scientific circle to the release of the Avatar film can be classified to any of the three classes: arts, 3D, and movies; an image showing a tiger in woods is associated with several keywords such as trees and tiger. In sum, one label per object is unable to fully describe such scenarios, and therefore, the research topic of multilabel classification has attracted increasing interest [31], [35], [39], [45], [48], [58], [62].

In multilabel learning, there are various challenges in multilabel data that would affect the learning process, such as label correlation [13] and [53], high dimensionality [28], [30], [32], streaming features [47], class imbalance [6], [7], [42], and label-specific features [58]. In this paper, we will focus on the first three challenges. First, different from the traditional single-label learning where the labels are mutually exclusive, the labels in multilabel learning are typically correlated and interdependent, which bring more difficulties to predict all relevant labels for a given object [22], [26]. For example, in automatic image annotation, "outdoor" and "sky" tend to appear in the same image; in text categorization, a document is relevant to multiple themes, such as "economy" and "sport"; in music information retrieval, a piece of symphony could convey various messages such as "piano," "classical music," and so on. On the other hand, it is a possibility to infer the unknown labels of an object from the known labels based on the label correlation. To fully utilize the relation between labels, Yu *et al.* [53] constructed a multilabel classification method based on the correlation between labels and the uncertainty between feature space and label space. Elisseeff and Weston [13] proposed to learn the ranks of labels for each instance, based on a large margin ranking system that shares a lot of common properties with support vector machines.

The second challenge is the high dimensionality of multilabel data, which usually has thousands or even tens of thousands of features. This is a common characteristic in image annotation and text categorization specially. For example, millions of informative words are extracted from a collection of documents or web pages to reflect their topics. As we know, some features are redundant and/or irrelevant for a given learning task, and high-dimensional data may bring several disadvantages to a learner [1], [8], [9], [25], [27], [33], [36], [46], [54]. To solve this problem, a number of multilabel dimensionality reduction approaches have been proposed. These methods can be divided into two groups: multilabel feature extraction and multilabel feature selection. Multilabel feature extraction converts the original high-dimensional feature space into a new low-dimensional

Manuscript received March 1, 2016; revised June 28, 2016 and May 10, 2017; accepted July 3, 2017. Date of publication August 3, 2017; date of current version November 29, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61672272, Grant 61303131, and Grant 61432011, in part by the China Postdoctoral Science Foundation (2015M581298), and in part by the US NSF IIS-1652107. *(Corresponding author: Qinghua Hu.)*

Y. Lin is with the School of Computer Science, Minnan Normal University, Zhangzhou 363000, China, and also with the School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: yjlin@mnnu.edu.cn).

Q. Hu is with the School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: huqinghua@tju.edu.cn).

J. Liu and J. Li are with the School of Computer Science, Minnan Normal University, Zhangzhou 363000, China (e-mail: zzliujinghua@163.com; jinjinli@mnnu.edu.cn).

X. Wu is with the School of Computing and Informatics, University of Louisiana at Lafayette, Louisiana, LA 70503 USA (e-mail: xwu@louisiana.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2017.2735947

feature space via mapping or transforming, however, it blurs the information of the original features and is lack of semantic interpretation. At present, several multilabel feature extraction techniques have been presented, such as multilabel dimensionality reduction via dependence maximization (MDDM) [59], linear discriminant analysis [23], canonical correlation analysis [15], and multilabel informed latent semantic indexing [52]. Different from multilabel feature extraction, multilabel feature selection ranks features based on the importance of each feature, and keeps the physical interpretation for selected features. Recently, a number of multilabel feature selection algorithms have been presented from different computing paradigms, such as information metric [30], [32], large margin [21], [41], and random forest [18].

The third challenge is that the full feature space is unknown in advance for some practical applications in various settings, i.e., feature dimensions continuously increase and not all features are available for learning while leaving the number of objects constant. This scenario is called streaming feature selection [47], [51]. For example, Mars crater detection [10] from high-resolution planetary images provides the only solution for remotely measuring the relative ages of planetary surfaces, and it is infeasible to generate and store tens of thousands of texture-based image features from planetary images to have a near global coverage of the Martian surface. Streaming features embrace a feature vector that flows in one by one over time while the number of training examples remains fixed, and the characteristics of multilabel feature selection in streaming features include: 1) feature dimensions may grow over time and may even extend to an infinite size; and 2) features flow in one at a time and each feature is required to be processed online upon its arrival. Therefore, we need to design an online feature selection method when features arrive one at a time. In addition, streaming features also exist in some real-world applications, such as email spam filtering [34], [50], and texture-based image segmentation [60]. To date, several research efforts have been made to address the challenge of streaming features, such as Grafting [37], Alpha-investing [61], Online Streaming Feature Selection (OSFS) [47], Online Group Feature Selection (OGFS) [60], and Scalable and Accurate OnLine Approach (SAOLA) [51]. However, these streaming feature selection algorithms are presented for single-label learning. To the best of our knowledge, no streaming feature selection algorithm for multilabel learning has been reported so far.

Indeed, a number of existing multilabel dimensionality reduction algorithms are effective in selecting an optimal feature subset for various multilabel learning tasks, but their solutions are only originated from one or two challenges mentioned previously, and they therefore cannot deal with these three challenges together, especially for streaming features. Motivated by these observations, we present two new multilabel feature selection algorithms to solve these challenges. First, we utilize fuzzy mutual information as an evaluation criterion of multilabel feature selection, and present a new data representation schema for categorical label spaces when computing the fuzzy mutual information between features and the whole label space, in which the data representation schema can capture the correlation between labels. Second, we present a standard multilabel feature selection algorithm that combines fuzzy mutual information with the Max-Relevance-Min-Redundancy strategy, and incorporates label correlation simultaneously. Finally, we propose a multilabel streaming feature selection (MSFS) algorithm for selecting features from streaming features, which is inspired by online relevance analysis and online redundancy analysis, and this algorithm can solve these three challenges in one shot. Experiments on various benchmark multilabel datasets show that the two proposed algorithms outperform existing state-of-the-art methods.

In summary, the contributions of this paper are three folds. First, we introduce a new data representation for nominal data, and present a measure of the correlation between labels. Second, two feature selection algorithms, i.e., multilabel feature selection with label correlation (MUCO) and MSFS, are designed for selecting static and streaming features, respectively. Finally, the effectiveness of the proposed algorithms is discussed with extensive experimental analysis.

The rest of this paper is organized as follows. Section II introduces multilabel learning and fuzzy mutual information. In Section III, we present two multilabel feature selection algorithms with label correlation for known and unknown feature spaces, respectively. Our experiments on benchmark datasets are demonstrated in Section IV. Finally, our conclusions and future work are given in Section V.

II. Preliminaries

A. Multilabel Learning

Let $T = (U, F, L)$ be a multilabel decision table, where $U = \{x_1, x_2, \ldots, x_n\}$ are n objects, $F = \{f_1, f_2, \ldots, f_m\}$ are m features, and $L = \{l_1, l_2, \ldots, l_k\}$ are k labels, respectively. Each object belongs to a subset of L and this subset can be described as a k-dimensional vector $y = [y^1, y^2, \ldots, y^k]$ where $y^j = 1$ only if x has label l_j; and 0 otherwise. The task of multilabel learning is to learn a function $h : U \to 2^L$.

At present, a number of multilabel learning algorithms have been proposed from different viewpoints, and these algorithms can be grouped into two categories: problem transformation and algorithm adaptation. The problem transformation approach addresses the multilabel learning problem via transforming it into other well-established learning scenarios, including binary classification, label ranking, and multilabel classification. On the other hand, the algorithm adaptation approach solves multilabel learning via adapting other effective learning techniques to deal with multilabel data directly, including decision tree construction, kernel learning, lazy learning, and information-theoretic induction. For analyzing different aspects of multilearning, some insightful reviews on multilabel learning are available in [20] and [57].

In multilabel experimental evaluation, we select some measures proposed in [40]. Let $T = \{(x_i, y_i)|1 \le i \le N\}$ be a test set where $y_i \subseteq L$ is a true label subset, and $\mathbf{y}_i^{'} \subseteq L$ be the binary label vector predicted by a learner for object x_i.

1) Average Precision (AP): This measure evaluates the average fraction of labels ranked higher than a particular label

$\gamma \in y_i$.

$$\text{AP} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|y_i|}\sum_{\gamma \in y_i}\frac{|\{\gamma^{'} \in y_i : r_i(\gamma^{'}) \leq r_i(\gamma)\}|}{r_i(\gamma)} \quad (1)$$

where $r_i(l)$ stands for the rank of label $l \in L$ predicted by the learner for x_i.

2) Coverage (CV): This measure evaluates how many steps are needed, on average, to go down the label ranking list so as to cover all the ground-truth labels of the object.

$$CV = \frac{1}{N}\sum_{i=1}^{N}\max_{\lambda \in y_i}\text{rank}(\lambda) - 1 \quad (2)$$

where $\text{rank}(\lambda)$ denotes the rank list of λ in terms of its likelihood. For example, if $\lambda_1 > \lambda_2$, then $\text{rank}(\lambda_1) < \text{rank}(\lambda_2)$.

3) One Error (OE): This measure evaluates the fraction of examples whose top-ranked label is not in the set of proper labels.

$$\text{OE} = \frac{1}{N}\sum_{i=1}^{N}[\![[\text{argmax}_{y_i \subseteq L} f(x_i, y_i)] \notin \mathbf{y}_i^{'}]\!] \quad (3)$$

where for any predicate π, $[\![\pi]\!]$ equals 1 if π holds and 0 otherwise.

4) Ranking Loss (RL): This measure evaluates the fraction of reversely ordered label pairs.

$$\text{RL} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|y_i||\overline{y_i}|}|\{(\lambda_1, \lambda_2)|\lambda_1 \leq \lambda_2, (\lambda_1, \lambda_2) \in y_i \times \overline{y_i}\}| \quad (4)$$

where λ_j is a real-valued likelihood between x_i and each label $l_i \in L$ based a classifier, and $\overline{y_i}$ denotes the complementary set of y_i.

5) Hamming Loss (HL): This measure evaluates how many times an instance-label pair is misclassified.

$$\text{HL} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathbf{y}_i^{'} \oplus y_i|}{m} \quad (5)$$

where $\oplus$ denotes the XOR operation.

6) Micro-F1 (F1): This measure calculates the F1 measure on the predictions of different labels as a whole.

$$F1 = \frac{2 \times \sum_{i=1}^{N}||\mathbf{y}_i^{'} \bigcap y_i||_1}{\sum_i^N ||y_i||_1 + \sum_i^N ||\mathbf{y}_i^{'}||_1} \quad (6)$$

where $\bigcap$ denotes the intersection operation, and $||\cdot||_1$ is the l_1-norm.

For these evaluation criteria, AP, CV, OE, and RL concern with the label ranking performance for each instance, and HL and F1 concern with the performance on label set prediction for each instance.

B. Fuzzy Entropy and Fuzzy Mutual Information

In this section, we introduce fuzzy entropy and fuzzy mutual information.

Given a nonempty finite set of objects $U = \{x_1, x_2, \ldots, x_n\}$ described by a set of features F, and R is a fuzzy equivalence relation over U generated by F, the fuzzy relation matrix $M(R)$ is defined as

$$M(R) = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}$$

where $r_{ij} \in [0, 1]$ is the relation value between x_i and x_j. Here, R satisfies the followings.

1) *Reflexivity:* $R(x, x) = 1\ \forall x \in U$.
2) *Symmetry:* $R(x, y) = R(y, x)\ \forall x, y \in U$.
3) *Transitivity:* $R(x, z) \geq \min_y\{R(x, y), R(y, z)\}$.

For all $x, y \in U$, some operations on relation matrices are defined as follows:

1) $R_1 = R_2 \Leftrightarrow R_1(x, y) = R_2(x, y)$;
2) $R = R_1 \bigcup R_2 \Leftrightarrow R(x, y) = \max\{R_1(x, y), R_2(x, y)\}$;
3) $R = R_1 \bigcap R_2 \Leftrightarrow R(x, y) = \min\{R_1(x, y), R_2(x, y)\}$;
4) $R_1 \subseteq R_2 \Leftrightarrow R_1(x, y) \leq R_2(x, y)$.

A fuzzy equivalence class associated with x_i and a fuzzy set R can be written as

$$[x_i]_R = [x_i]_F = r_{i1}/x_1 + r_{i2}/x_2 + \cdots + r_{in}/x_n$$

where r_{ij} is the degree of x_i equivalent to x_j, "+" means "union," and the symbol "/" means a separator.

Then, the fuzzy cardinality of the fuzzy equivalence class is defined as

$$|[x_i]_R| = |[x_i]_F| = \sum\nolimits_{j=1}^{n} r_{ij}.$$

Definition 1: [24]. Given an approximate space $< U, R >$, the fuzzy information entropy of F is defined as

$$\text{FH}(R) = \text{FH}(F) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_R|}{n} \quad (7)$$

where $|[x_i]_R| = \sum_{j=1}^{n} r_{ij}$.

If R is a crisp equivalence relation, namely $r_{ij} \in \{0, 1\}$, then the fuzzy information entropy can be degenerated into Shannon's information entropy.

Definition 2: [24]. Let F_1 and F_2 be two subsets of F, then the fuzzy joint entropy is computed as

$$\text{FH}(F_1, F_2) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1} \bigcap [x_i]_{F_2}|}{n} \quad (8)$$

where $[x_i]_{F_1} \bigcap [x_i]_{F_2} = \min\{[x_i]_{F_1}, [x_i]_{F_2}\}$.

Definition 3: [24]. Let F_1 and F_2 be two subsets of F. The fuzzy conditional entropy of F_2 conditioned to F_1 is defined as

$$\text{FH}(F_2|F_1) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1} \bigcap [x_i]_{F_2}|}{|[x_i]_{F_1}|}. \quad (9)$$

Theorem 1: $\text{FH}(F_2|F_1) = \text{FH}(F_1, F_2) - \text{FH}(F_1)$.

Theorem 2: Let F_1 and F_2 be two subsets of F. Then, we have

1) $\text{FH}(F_1, F_2) \geq \max\{\text{FH}(F_1), \text{FH}(F_2)\}$;
2) $F_1 \subseteq F_2$ or $R_{F_1} \subseteq R_{F_2} : \text{FH}(F_1, F_2) = \text{FH}(F_1)$;
3) $F_1 \subseteq F_2$ or $R_{F_1} \subseteq R_{F_2} : \text{FH}(F_2|F_1) = 0$.

Definition 4: [24]. Let F_1 and F_2 be two subsets of F. Then, the fuzzy mutual information between F_1 and F_2 is defined as

$$\text{FMI}(F_1; F_2) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1}| \cdot |[x_i]_{F_2}|}{n \cdot |[x_i]_{F_1} \bigcap [x_i]_{F_2}|}. \quad (10)$$

Theorem 3: $\text{FMI}(F_1; F_2) = \text{FH}(F_1) - \text{FH}(F_1|F_2) = \text{FH}(F_2) - \text{FH}(F_2|F_1)$.

From (7)–(10), fuzzy entropy and fuzzy mutual information can be used to compute hybrid data, and it overcomes the hybrid data limitation of Shannon's mutual information.

III. Proposed Algorithms

A. Problem Statement

Given a feature space F on a training dataset, feature selection is to select a compact feature subset from F without performance degradation for prediction models. Based on the information theory, Bell *et al.* [2] introduced a feature selection mechanism according to the first axiomatic method.

Axiom 1.1 (Preservation of learning information): For a given dataset described by features F and a decision variable C, if there exists a feature subset S such that $I(S; C) = I(F; C)$, then S is sufficient, where $I(S; C)$ denotes the mutual information between S and C.

Axiom 1.2 (Minimum encoding length): Suppose a given dataset described by features F and a decision variable C, and $\mathcal{S}$ is a feature subset. Every $s \in \mathcal{S}$, which minimizes the joint entropy $H(s, C)$, should be favored with respect to its predictive ability.

Axioms 1.1 and 1.2 provide an axiomatic description of a good feature subset based on the information theory and the principle of Occams razor. Similarly, we can give the second axiomatic approach that is suitable for MUCO.

Axiom 2.1 (Preservation of learning information for multilabel feature selection): Given a feature space F and a label space L in a multilabel decision table, if there exists $S \subseteq F$ such that $\text{FMI}(S; L) = \text{FMI}(F; L)$, then S is sufficient with respect to the multilabel decision table.

Axiom 2.2 (Minimum encoding length for multilabel feature selection): Given a feature space F, a label space L in a multilabel decision table, and a set of sufficient feature subset $\mathcal{S}$, for any $s \in \mathcal{S}$, which minimizes the joint entropy $\text{FH}(s, L)$, should be favored with respect to its predictive ability for the multilabel decision table.

Axioms 1.1 and 1.2 provide a criterion for single-label feature selection, and Axioms 2.1 and 2.2 provide a criterion for multilabel feature selection from the information theory, respectively. According to Axioms 1.1 and 1.2, Yu *et al.* [49] used fuzzy mutual information to pursue heterogeneous feature selection based on min-Redundancy-Max-Relevance, Max-Dependency, and min-Redundancy-Max-Dependency, respectively. In addition, Peng *et al.* [36] proposed a popular single-label feature selection method based on redundancy analysis and relevance analysis, called minimal-redundancy and max-relevance (mRMR). Based on Axioms 2.1 and 2.2, Lin *et al.* [33] utilized conditional redundancy analysis and dependence analysis to do multilabel feature selection. Different from [33], [36], and [49], in this paper, the novelties of our proposed methods include: 1) MUCO extends the strategy of redundancy analysis and relevance analysis to do multilabel feature selection based on fuzzy mutual information; 2) MSFS addresses streaming feature selection for multilabel learning; and 3) both MUCO and MSFS incorporate label correlation.

According to the aforementioned discussion, in this paper, we can first give the optimization objective function as follows for standard multilabel feature selection.

$$S = \underset{S}{\text{argmin}}\{|S| : S = \underset{\varsigma \subseteq F}{\text{argmax}}\,\text{FMI}(\varsigma; L)\} \quad (11)$$

where F denotes the whole known feature space, L is the label space, and $S \subseteq F$ is the final selected feature subset.

Equation (11) gives a criterion to do multilabel feature selection when the feature space is known, which includes two respects: the discriminative ability of the selected features should be more than or at least equal to the original feature space, and the number of the selected features should be as small as possible.

Different from the standard multilabel feature selection, which is formulated by (11), the challenge of streaming feature selection is that, as we process one feature at a time, how to online obtain a subset of informative features S'_{t_i} to maximize its predictive performance at any time t_i. Therefore, we can present the optimization objective function as (12) for MSFS.

$$S'_{t_i} = \underset{S'}{\text{argmin}}\{|S'| : S' = \underset{\varsigma \subseteq (S'_{t_{i-1}} \bigcup \{f_i\})}{\text{argmax}}\,\text{FMI}(\varsigma; L)\} \quad (12)$$

where f_i is a new arriving feature at time t_i, and $S'_{t_{i-1}} \subseteq F$ is the selected feature subset at time t_{i-1}.

Different from (11), (12) gives another criterion to do MSFS when the feature space is unknown, which also includes two respects: the distinguishing ability of the selected features should be maximal for all arrived features at any specified time, and the number of the selected features should be as small as possible.

B. Label Correlation

To capture the correlation between labels, we consider all other labels' influences on each label by building a similarity matrix for data objects under the whole label space. In (11) and (12), a key factor of multilabel feature selection is to compute $\text{FMI}(\varsigma; L)$. Therefore, we need to construct a fuzzy relation matrix $M(\bar{R})$ for objects under the label space L. To describe the correlation between labels precisely, we present a data representation scheme for categorical label data, which maps a set of categorical values of labels into a Euclidean space, then a much finer-grained metric for measuring the similarity between objects under the label space is proposed. The advantage of this transformation is described in [38].

Let $T = (U, F, L)$ be a multilabel decision table, where $U = \{x_1, x_2, \ldots, x_n\}$ are n objects, $F = \{f_1, f_2, \ldots, f_m\}$ are m features, and $L = \{l_1, l_2, \ldots, l_k\}$ are k labels, where the $l_i(x_j)$ is a categorical value to denote whether object x_j has the label l_i. Now, we give a new data representation scheme for categorical label data. First, we use the Jaccard coefficient to measure the

similarity degree between objects under L, which is formally defined as

$$c_{ij} = \frac{|L(x_i) \bigcap L(x_j)|}{|L(x_i) \bigcup L(x_j)|} \tag{13}$$

where $L(x_i)$ represents the label set of x_i.

Equation (13) maps a set of categorical label data into a Euclidean space. However, the Jaccard coefficient ignores the difference between objects without the same labels, and the number of positive objects with respect to each class label is far less than its negative counterparts in multilabel learning. Therefore, we need to redefine a new metric to measure the similarity between objects in the Euclidean space as follows:

$$r_{ij}^{L} = 1 - \frac{d(x_i, x_j)}{\max(d(x_s, x_t)) - \min(d(x_s, x_t))} \tag{14}$$

where $s, t = 1, 2, \ldots, n$, and $d(x_i, x_j) = (\sum_{r=1}^{n}(c_{ir} - c_{jr})^2)^{\frac{1}{2}}$.

By (14), we can obtain the fuzzy relation matrix $M(\bar{R})$ between objects under the label space L. Meanwhile, $M(\bar{R})$ reflects the inner correlation between labels.

C. MUCO

From Axioms 2.1 and 2.2, we know that a straightforward way to select an expected feature subset is to exhaustively evaluate all features. However, this is not practical even given a medium size of candidate features according to the exponential complexity. Therefore, some efficient algorithms were designed to overcome this problem, where the mRMR [36] is a popular criterion, and has proven its efficiency and effectiveness. In this section, we use the Max-Relevance and Min-Redundancy (MRMR) strategy for multilabel feature selection, i.e., a candidate feature is selected if it is totally relevant to all labels, and is not redundant with the selected features.

The first step of the MRMR for multilabel feature selection is to select a candidate feature that has the maximal relevance with the label set, referred to as Max-Relevance. Let S be the selected features and L be the label set. The Max-Relevance is formulated as

$$\max D(S, L), \text{where } D(S, L) = \frac{1}{|S|} \sum_{f_i \in S} \text{FMI}(f_i; L). \tag{15}$$

However, multilabel feature selection with Max-Relevance may include redundancy, i.e., the new selected feature f_i is strongly relevant to some features selected previously. Therefore, we should measure the redundancy between a candidate feature and the selected features in the process of feature selection, then a Min-Redundancy metric is defined as

$$\min R(S), \text{where } R(S) = \frac{1}{|S|^2} \sum_{f_i, f_j \in S} \text{FMI}(f_i; f_j). \tag{16}$$

The metric combining the aforementioned two constraints is called Max-Relevance and Min-Redundancy. Then, we define an operator $\Phi(D, R)$ to combine D and R, and optimize D and R via $\Phi(D, R)$, simultaneously

$$\max \Phi(D, R), \Phi(D, R) = D(S, L) - R(S). \tag{17}$$

Algorithm 1: MUCO.

Input: F: A set of candidate features; U: A set of objects; L: A set of labels, f: A candidate feature.
Output: $\mathbf{s}$: the feature vector: $\mathbf{s} = (s_1, s_2, \ldots, s_{|F|})$.

```
Initialize s = [], k = 1;
while |F| ≠ ∅ do
    find f ∈ F by maximizing (18);
    s_k = f;
    F = F − {f};
    k = k + 1;
end while
return s.
```

Given the set S_{k-1} with $k-1$ features selected, the kth feature can be determined by

$$\max_{f_j \in F - S_{k-1}} [\text{FMI}(f_j; L) - \frac{1}{k-1} \sum_{f_i \in S_{k-1}} (\text{FMI}(f_i; f_j))]. \tag{18}$$

In the aforementioned equation, the first term analyzes the relevance between the candidate feature f_j and the label set L, and the second term focuses on the redundancy between the candidate feature f_j and the selected features S_{k-1}. Therefore, based on MRMR and label correlation, we propose a multilabel feature selection algorithm as Algorithm 1.

There are two key steps in Algorithm 1. The first step is to construct a fuzzy relation matrix under the feature space and the label space, respectively, and their time complexities are both $O(|U| \cdot |U|)$. The other step is an incremental search, and its time complexity is $O(|S| \cdot |F|)$, where $|S|$ is the number of selected features , and $|F|$ is the number of candidate features. Therefore, the computational complexity of this algorithm is $O(|U| \cdot |U| + |S| \cdot |F|)$.

D. MSFS With Label Correlation

MUCO assumes that all candidate features are available for a learner before feature selection takes place. In contrast, MSFS assumes that the size of the feature set is unknown, and not all features are available for learning while leaving the number of objects constant. Based on Axioms 2.1 and 2.2, we present an MSFS method with two steps: online relevance analysis and online redundancy analysis.

1) *Online Relevance Analysis:* Assume $S'_{t_{i-1}}$ is the selected feature subset at time t_{i-1}, and a new feature f_i comes at time t_i. Given a relevance threshold δ, if $\text{FMI}(f_i; L) \geq \delta (0 < \delta < 1)$, f_i is said to be a relevant feature to L; otherwise, f_i is discarded as an irrelevant feature and will never be considered again.
2) *Online Redundancy Analysis:* Assume $S'_{t_{i-1}}$ is the selected feature subset at time t_{i-1}, and a relevant feature f_i comes at time t_i. If there exists $f_s \in S'_{t_{i-1}}$ such that $\text{FMI}(f_i; L | f_s) = 0$, it testifies that adding f_i alone to $S'_{t_{i-1}}$ does not increase the predictive capability of $S'_{t_{i-1}}$. Based on this observation, we give the following lemma.

Algorithm 2: MSFS.

Input: L: A set of labels; f_i: Features arrive at time t_i; δ: A relevance threshold ($0 \leq \delta \leq 1$); $S'_{t_{i-1}}$: The selected features at time t_{i-1}.
Output: S'_{t_i}: The selected features at time t_i.

1: Build fuzzy relation matrix $M(R_L)$ on label space L and fuzzy equivalence relation matrix $M(R_f)$ on feature f;
2: **repeat**
3: a new feature f_i arrives at time t_i;
4: **if** FMI$(f_i; L) \leq \delta$ **then**
5: discard f_i and go to step 16;
6: **end if**
7: **for** each feature $f_s \in S'_{t_{i-1}}$ **do**
8: **if** FMI$(f_i; L) <$ FMI$(f_s; L)$ **then**
9: discard f_i and go to step 16;
10: **end if**
11: **if** FMI$(f_i; L) >$ FMI$(f_s; L)$ **then**
12: $S'_{t_{i-1}} = S'_{t_{i-1}} - \{f_s\}$;
13: **end if**
14: **end for**
15: $S'_{t_i} = S'_{t_{i-1}} \bigcup \{f_i\}$.
16: **until** no features are available

Lemma 1: With the current feature subset $S'_{t_{i-1}}$ at time t_{i-1} and a new feature f_i at time t_i, if there exists $f_s \in S'_{t_{i-1}}$ such that $FMI(f_i; L|f_s) = 0$, then $FMI(f_s; L) > FMI(f_i; L)$.

Proof: As FMI$(f_i; f_s|L)$ − FMI$(f_i; f_s)$ = FMI$(f_i; L|f_s)$ − FMI$(f_i; L)$. If FMI$(f_i; L|f_s) = 0$, then

$$\text{FMI}(f_i; f_s) = \text{FMI}(f_i; f_s|L) + \text{FMI}(f_i; L). \quad (19)$$

■

Making use of the identity FMI$(f_s; L|f_i) =$ FMI$(f_s; L) +$ FMI$(f_s; f_i|L) -$ FMI$(f_i; f_s)$ and (19), we can get (20) as follows:

$$\text{FMI}(f_s; L|f_i) = \text{FMI}(f_s; L) - \text{FMI}(f_i; L). \quad (20)$$

Since f_s is in the current feature set $S'_{t_{i-1}}$, FMI$(f_s; L|f_i) > 0$. Accordingly, the following holds.

$$\text{FMI}(f_s; L) > \text{FMI}(f_i; L). \quad \square \quad (21)$$

Based on the aforementioned analysis, we can get Proposition 1 to select or discard the feature f_i.

Proposition 1: With the current feature subset $S'_{t_{i-1}}$ at time t_{i-1} and a new feature f_i at time t_i, if $FMI(f_i; L) < \delta$, then f_i is discarded. Moreover, if there exists $f_s \in S'_{t_{i-1}}$ such that $FMI(f_i; L) \geq \delta$ and $FMI(f_i; L) < FMI(f_s; L)$, then f_i is discarded. Otherwise, f_i is added to $S'_{t_{i-1}}$.

With the combination of online relevance analysis and online redundancy analysis, the algorithm of the MSFS with label correlation (MSFS for short) can be formed in Algorithm 2.

The major computation in MSFS is to compute the correlations between features. At time t_i, assuming S'_{t_i} is the number of the currently selected features, and the total number of features is up to Q, then the time complexity of the MSFS is $O(|Q| \cdot |S'_{t_i}|)$.

TABLE I
CHARACTERISTICS OF BENCHMARK MULTILABEL DATASETS

Dataset	Instances	Features	Labels	Training	Test	Card	Density
Arts	5000	462	26	2000	3000	1.636	0.063
Birds	645	260	19	322	323	1.470	0.074
Business	5000	438	30	2000	3000	1.588	0.053
Cal500	502	68	174	251	251	26.044	0.150
Computer	5000	681	33	2000	3000	1.508	0.046
Education	5000	550	33	2000	3000	1.461	0.044
Emotions	593	72	6	391	202	1.869	0.311
Health	5000	612	32	2000	3000	1.662	0.052
Recreation	5000	606	22	2000	3000	1.423	0.065
Reference	5000	793	33	2000	3000	1.169	0.035
Society	5000	636	27	2000	3000	1.692	0.063
Yeast	2417	103	14	1499	918	4.238	0.303

IV. EXPERIMENTAL ANALYSIS

In this section, we empirically demonstrate the superiority of the proposed algorithms. We first present the characteristics of our datasets, comparative methods, evaluation metrics, and base classifier, respectively. Then, the performance analysis on MUCO and MSFS is reported. Finally, we analyze the efficiency between MUCO and MSFS.

A. Datasets and Experimental Settings

1) Datasets: To test MUCO, we select 12 benchmark datasets from different application domains in Mulan Library [44], [59]. Among these datasets, Arts, Business, Computer, Health, Recreation, Reference, and Society are frequently used to text categorization. CAL500 is composed of 500 popular western musical tracks. Birds is a real-world dataset of bird sounds collected in field conditions, which consists of 645 ten-second audio recordings in uncompressed WAV format, and there are 19 species of birds. Emotions is a benchmark for music, which contains 593 music objects and each object belongs to a subset of six labels. Yeast has 2 417 objects where each object represents a yeast gene and there are 14 labels indicating gene functional groups. Table I displays these characteristics of these multilabel datasets.

To verify the effectiveness of the MSFS, we select eight relatively high-dimensional multilabel datasets from Mulan Library [44], [59]. To fully illustrate the relationship between the MUCO and MSFS, we choose six datasets whose feature dimensions are higher than 500 from Table I, i.e., Computer, Education, Health, Recreation, Reference, and Society. In addition, we add other two high-dimensional multilabel datasets, i.e., Enron and Science. A summary of these eight datasets is given in Table II.

2) Experimental Settings: To illustrate the effectiveness of our proposed algorithms, we select five state-of-the-art multilabel feature selection methods as baselines, including Multilabel Dimensionality reduction via Dependence Maximization for uncorrelated projection dimensionality reduction (MDDMspc) [59], Multi-label Dimensionality reduction via Dependence Maximization for uncorrelated subspace dimensionality reduction (MDDMproj) [59], Pairwise Multivariate Mutual Information (PMU) [29], ReliefF for Multi-Label feature selection

TABLE II
SUMMARY OF HIGH-DIMENSIONAL MULTILABEL DATASETS WITH STREAMING FEATURES

Dataset	Instances	Features	Labels	Training	Test	Card	Density
Computer	5000	681	33	2000	3000	1.508	0.046
Education	5000	550	33	2000	3000	1.461	0.044
Enron	1702	1001	53	1123	579	1.953	0.037
Health	5000	612	32	2000	3000	1.662	0.052
Recreation	5000	606	22	2000	3000	1.423	0.065
Reference	5000	793	33	2000	3000	1.169	0.035
Science	5000	743	40	2000	3000	1.451	0.037
Society	5000	636	27	2000	3000	1.692	0.063

(RF-ML) [41], and feature selection for Multi-Label Naive Bayes classification (MLNB) [56]. In MDDMspc, μ is set as 0.5. In PMU, continuous features are discretized into two bins, and the categorical features are left untouched, as recommend in [29]. In MSFS, the parameter δ is set to 0.08. Meanwhile, the classification performance of all feature selection algorithms are evaluated using MLKNN ($K = 10$) [55]. Finally, we select *AP*, *CV*, *OE*, *RL*, *HL*, and *F1* as criteria to evaluate the performance of feature selection. Note that the six criteria were originated from different evaluation viewpoints, and usually few algorithms could outperform other algorithms on all these criteria.

B. Performance Analysis on MUCO

In this section, we group experiments by three aspects. In the first aspect, we compare MUCO with MDDMspc, MDDMproj, RF-ML, and PMU, as these algorithms obtain a feature rank list as their results of feature selection. In the second aspect, we compare the classification performance among MDDMspc, MDDMproj, PMU, MLNB, RF-ML, and MUCO, in which the number of selected features of other algorithms is equal to MLNB, for an impartial comparison. In the third aspect, we perform performance analysis based on statistical analysis among the comparing algorithms in a systematical way. In addition, similar to the five comparative feature selection algorithms, and for an impartial comparison, we also use the training set and test set as they have been already separated in Mulan Library.

To compare MUCO with MDDMspc, MDDMproj, RF-ML, and PMU, we conduct a number of experiments to demonstrate the change tendency of classification performance as the number of the selected features increases. Figs. 1–6 display the classification performance with all datasets on different evaluation metrics. In these figures, the horizontal axis represents the size of the selected features, and the vertical axis indicates the classification performance of different measures after feature selection, respectively. There are five lines in each figure, corresponding to MDDMspc, MDDMproj, RF-ML, PMU, and MUCO, respectively. For all criteria, the results of all these datasets are showed as variation tendency of different algorithms growing with the number of selected features. Finally, Figs. 1–6 show that MUCO obtains superior classification performance compared to other five popular feature selection methods on all datasets with different criteria.

TABLE III
COMPARISON BETWEEN MUCO AND OTHER FIVE ALGORITHMS IN TERMS OF AP (↑)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MUCO
Arts	0.5072	0.4943	0.4823	0.4944	0.4991	**0.5192**
Birds	0.6949	0.6949	0.6949	0.6949	0.5996	**0.6949**
Business	0.8736	0.8732	0.8739	0.8754	0.8713	**0.8770**
Cal500	0.4823	0.4823	0.4788	0.4796	0.4776	**0.4828**
Computer	0.6345	0.6284	0.6285	0.6276	0.6391	**0.6403**
Education	0.5389	0.5425	0.5365	0.5465	0.5478	**0.5753**
Emotions	0.7734	0.7627	0.7532	0.7081	0.7529	**0.7755**
Health	0.6654	0.6502	0.6686	0.6802	**0.6880**	0.6857
Recreation	0.4717	0.4703	0.4465	0.4365	**0.4790**	0.4775
Reference	0.6126	0.6106	0.6151	0.6169	0.6234	**0.6301**
Society	0.5615	0.5681	0.5900	0.5881	0.5894	**0.5939**
Yeast	0.7213	0.7210	0.7471	**0.7473**	0.7355	0.7350
Average	*0.6281*	*0.6249*	*0.6263*	*0.6246*	*0.6252*	**0.6406**

TABLE IV
COMPARISON BETWEEN MUCO AND OTHER FIVE ALGORITHMS IN TERMS OF RL (↓)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MUCO
Arts	0.1521	0.1555	0.1540	0.1527	0.1542	**0.1504**
Birds	0.1251	0.1251	0.1251	0.1251	0.1744	**0.1251**
Business	0.0422	0.0422	0.0417	0.0413	0.0419	**0.0403**
Cal500	0.1908	0.1908	0.1904	**0.1891**	0.1913	0.1905
Computer	0.0916	0.0934	0.0931	0.0941	0.0910	**0.0896**
Education	0.0914	0.0924	0.0939	0.0911	0.0922	**0.0856**
Emotions	**0.1762**	0.1940	0.2028	0.2656	0.2055	0.1850
Health	0.0663	0.0698	0.0643	0.0638	0.0641	**0.0608**
Recreation	0.1838	0.1859	0.1917	0.1955	0.1879	**0.1857**
Reference	0.0888	0.0889	**0.0856**	0.0868	0.0889	0.0865
Society	0.1500	0.1484	0.1443	0.1442	0.1456	**0.1416**
Yeast	0.1990	0.2041	0.1815	**0.1786**	0.1871	0.1909
Average	*0.1298*	*0.1325*	*0.1307*	*0.1357*	*0.1353*	**0.1277**

To demonstrate the effectiveness of MUCO more specifically and clearly, we select the top P features as the selected features, and P is determined by MLNB. For example, the number of the selected features with MLNB in Health is 303, and then, we set $P = 303$ for all algorithms on Health. Tables III–VIII show the classification performance obtained with MLNB, MDDMspc, MDDMproj, PMU, RF-ML, and MUCO, respectively. In these tables, bold font indicates the best performance for each dataset, italics indicates the average classification performance for each algorithm on all datasets, "↑" indicates "the larger the better," and "↓" denotes "the smaller the better," respectively. For all experimental results in these tables, we can observe that

1) MUCO totally outperforms MDDMspc, MDDMproj, PMU, RF-ML, and MLNB with all evaluation metrics;
2) MUCO achieves better performance against comparing algorithms on at least nine datasets with AP and RL, respectively. Note that the performance of the MUCO gets suboptimal on the other two datasets with AP and RL;
3) for CV, HL, and OE, MUCO performs better than comparing algorithms on at least seven datasets, and the performance of the MUCO is extremely close with the best value on other five datasets;

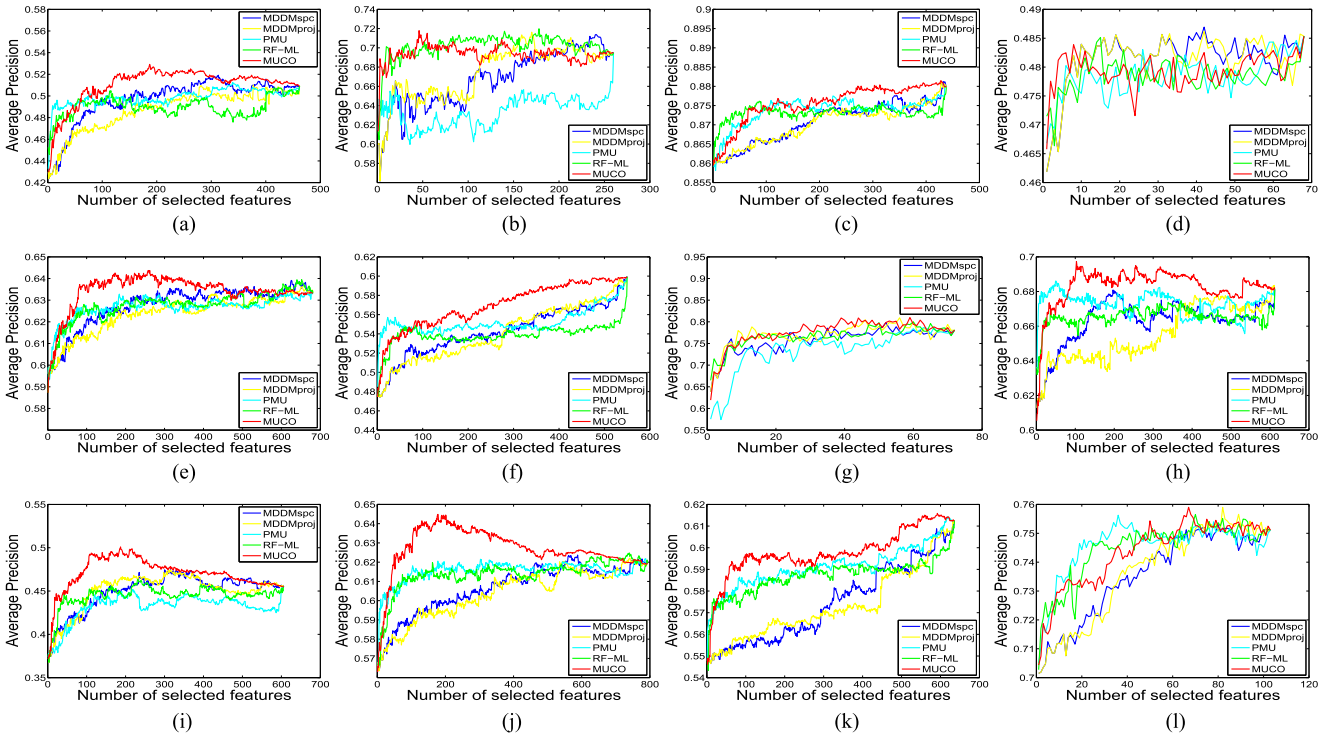

Fig. 1. Performance variation of selected features with respect to AP. (a) Arts. (b) Birds. (c) Business. (d) Cal500. (e) Computer. (f) Education. (g) Emotions. (h) Health. (i) Recreation. (j) Reference. (k) Society. (l) Yeast.

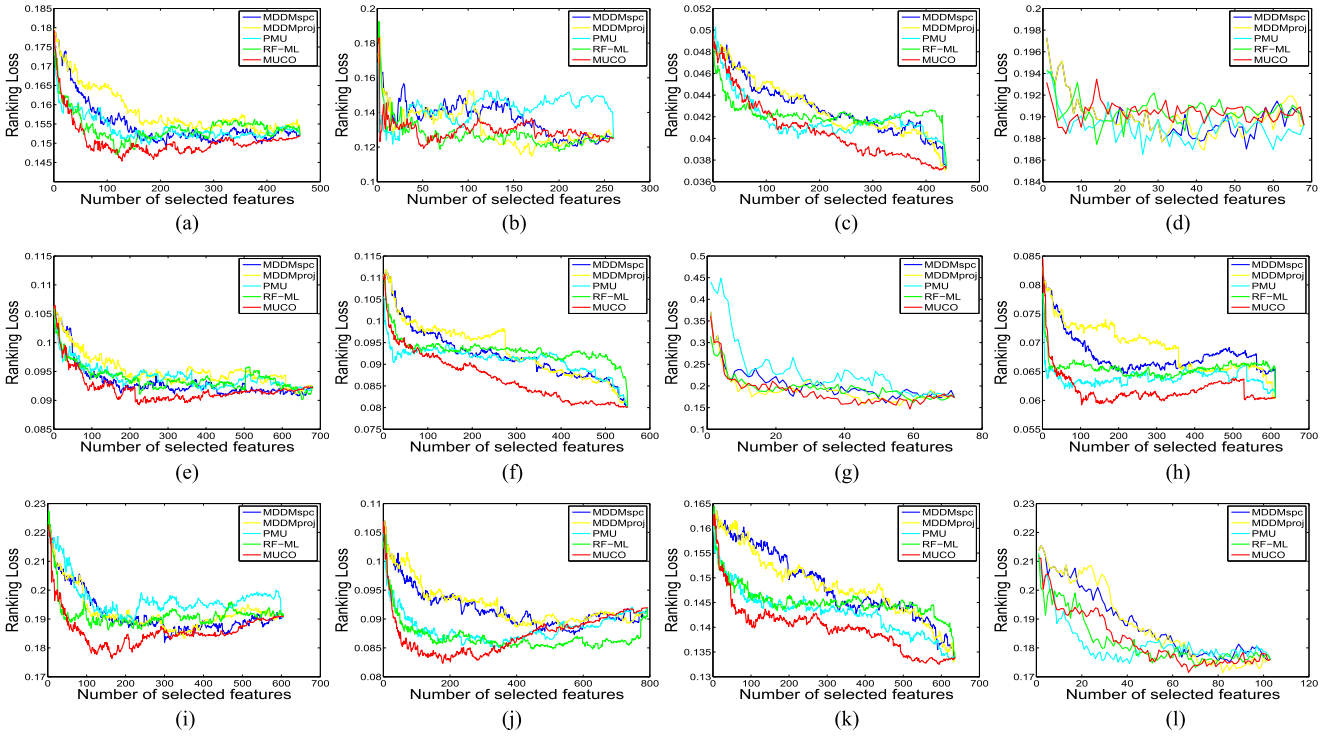

Fig. 2. Performance variation of selected features with respect to RL. (a) Arts. (b) Birds. (c) Business. (d) Cal500. (e) Computer. (f) Education. (g) Emotions. (h) Health. (i) Recreation. (j) Reference. (k) Society. (l) Yeast.

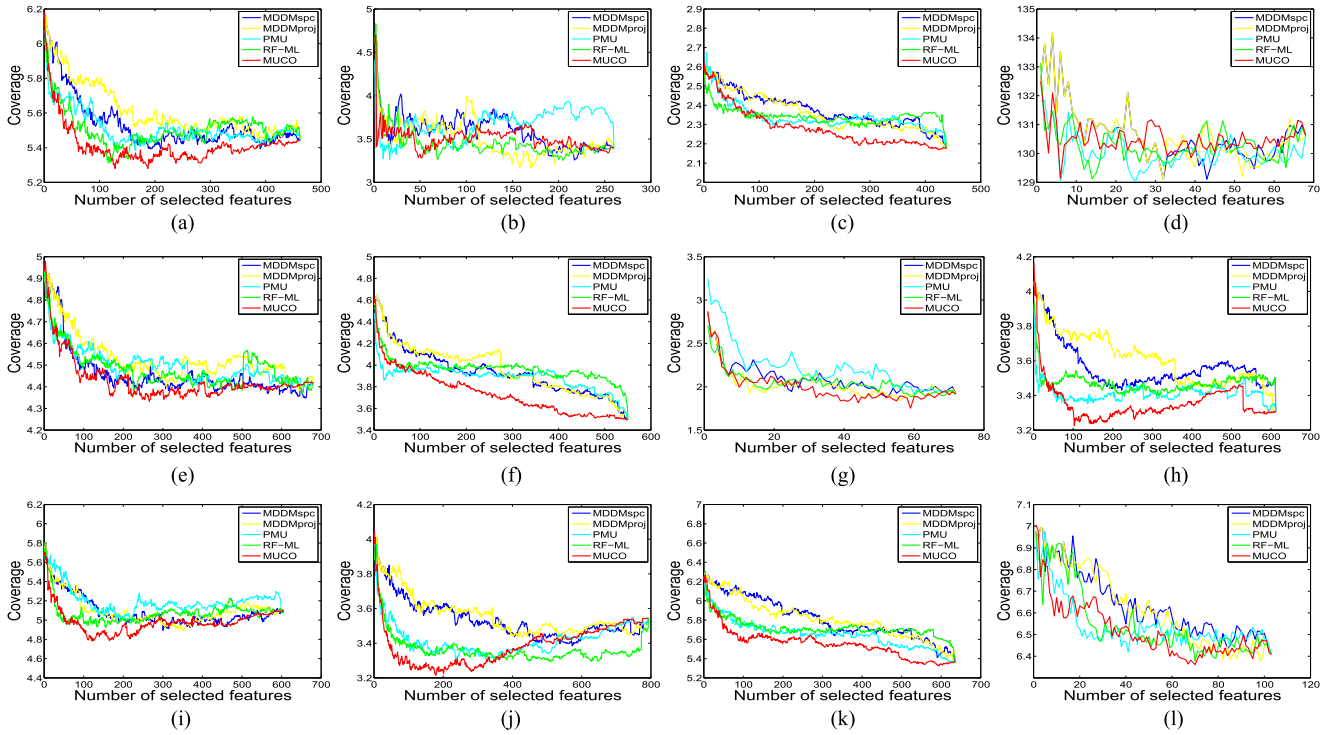

Fig. 3. Performance variation of selected features with respect to CV. (a) Arts. (b) Birds. (c) Business. (d) Cal500. (e) Computer. (f) Education. (g) Emotions. (h) Health. (i) Recreation. (j) Reference. (k) Society. (l) Yeast.

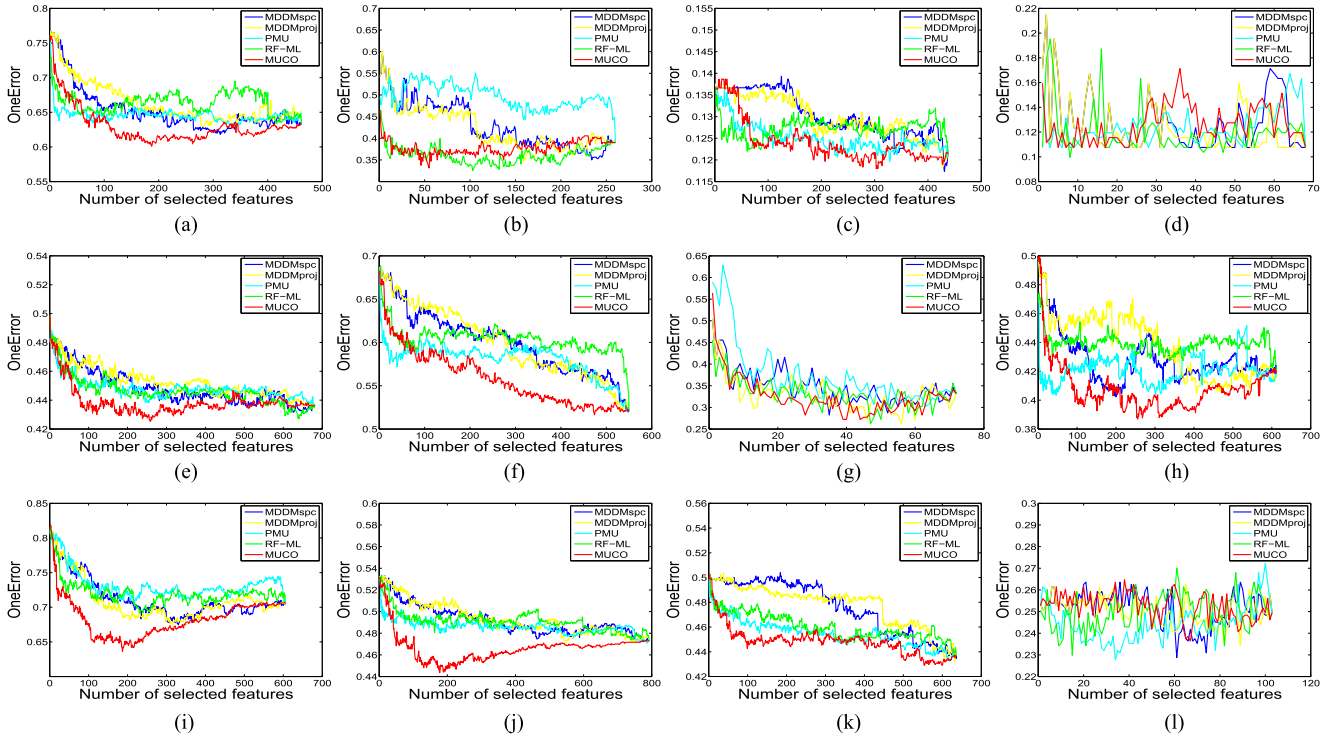

Fig. 4. Performance variation of selected features with respect to OE. (a) Arts. (b) Birds. (c) Business. (d) Cal500. (e) Computer. (f) Education. (g) Emotions. (h) Health. (i) Recreation. (j) Reference. (k) Society. (l) Yeast.

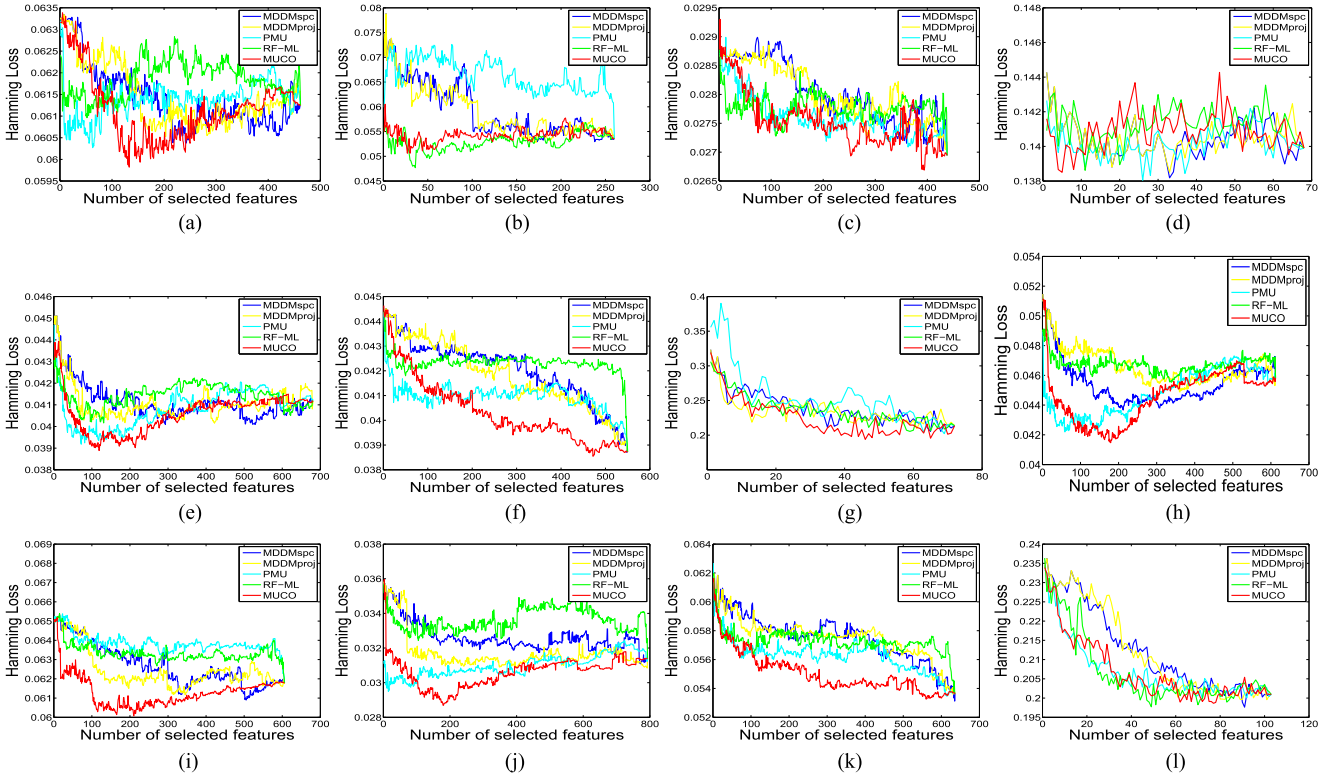

Fig. 5. Performance variation of selected features with respect to HL. (a) Arts. (b) Birds. (c) Business. (d) Cal500. (e) Computer. (f) Education. (g) Emotions. (h) Health. (i) Recreation. (j) Reference. (k) Society. (l) Yeast.

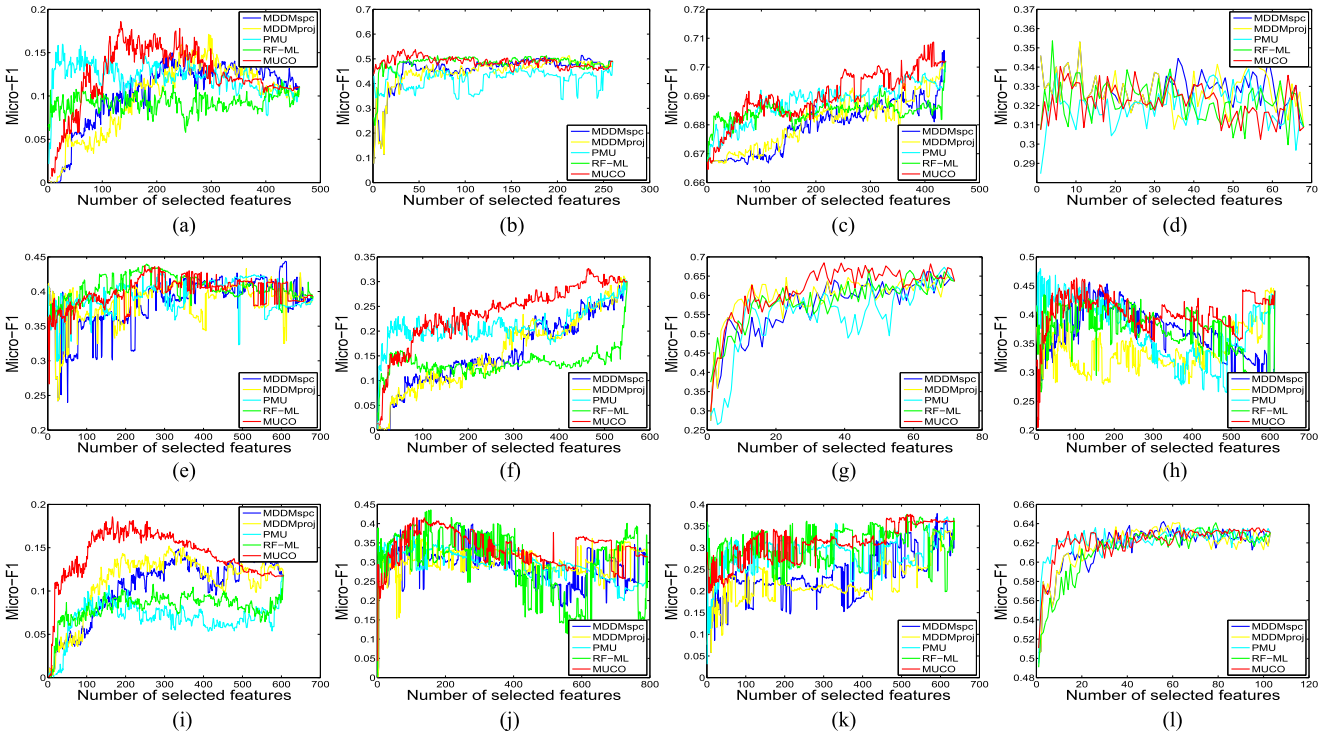

Fig. 6. Performance variation of selected features with respect to F1. (a) Arts. (b) Birds. (c) Business. (d) Cal500. (e) Computer. (f) Education. (g) Emotions. (h) Health. (i) Recreation. (j) Reference. (k) Society. (l) Yeast.

TABLE V
COMPARISON BETWEEN MUCO AND OTHER FIVE ALGORITHMS IN TERMS OF CV (↓)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MUCO
Arts	5.4740	5.5553	5.4853	5.4917	5.5040	**5.4187**
Birds	3.3994	3.3994	3.3994	3.3994	4.3653	**3.3994**
Business	2.3460	2.3303	2.3147	2.3187	2.3483	**2.2687**
Cal500	130.3506	130.3506	**129.4462**	129.8008	131.4343	130.2629
Computer	4.3987	4.4437	4.4427	4.5013	4.3740	**4.3487**
Education	3.8987	3.9203	3.9920	3.8990	3.9183	**3.6930**
Emotions	**1.9455**	2.0495	2.0792	2.4059	2.0743	2.0149
Health	3.5057	3.6217	3.4257	3.4070	3.4163	**3.3153**
Recreation	**4.9403**	4.9470	5.0860	5.1367	4.9953	5.0033
Reference	3.4390	3.4460	**3.3270**	3.3660	3.4313	3.3580
Society	5.8423	5.8000	5.6740	5.6603	5.7390	**5.6107**
Yeast	**6.8137**	6.8181	6.4771	6.4913	6.6928	6.6057
Average	*14.6962*	*14.7235*	**14.5958**	*14.6565*	*14.8578*	*14.6083*

TABLE VI
COMPARISON BETWEEN MUCO AND OTHER FIVE ALGORITHMS IN TERMS OF OE (↓)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MUCO
Arts	0.6340	0.6487	0.6790	0.6537	0.6433	**0.6137**
Birds	0.3901	0.3901	0.3901	0.3901	0.5418	**0.3901**
Business	0.1287	0.1280	0.1273	0.1227	0.1317	**0.1207**
Cal500	0.1474	0.1474	0.1195	0.1195	0.1434	**0.1076**
Computer	0.4403	0.4490	0.4483	0.4467	**0.4320**	0.4330
Education	0.6100	0.5973	0.6100	0.5920	0.5827	**0.5520**
Emotions	0.3218	0.3515	0.3614	0.3911	0.3762	**0.3168**
Health	0.4270	0.4403	0.4370	0.4080	**0.3947**	0.4013
Recreation	0.6793	0.6827	0.7113	0.7210	**0.6643**	0.6670
Reference	0.4843	0.4887	0.4930	0.4867	0.4703	**0.4637**
Society	0.4953	0.4813	0.4593	0.4593	0.4540	**0.4533**
Yeast	0.2593	0.2527	0.2397	**0.2331**	0.2560	0.2527
Average	*0.4181*	*0.4215*	*0.4230*	*0.4187*	*0.4242*	**0.3977**

TABLE VII
COMPARISON BETWEEN MUCO AND OTHER FIVE ALGORITHMS IN TERMS OF HL (↓)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MUCO
Arts	0.0607	0.0612	0.0627	0.0615	0.0612	**0.0605**
Birds	0.0536	0.0536	0.0536	0.0536	0.0748	**0.0536**
Business	0.0277	0.0277	0.0278	0.0273	0.0283	**0.0273**
Cal500	0.1396	0.1396	0.1399	0.1399	0.1426	**0.1395**
Computer	0.0406	0.0406	0.0521	0.0413	**0.0401**	0.0407
Education	0.0426	0.0422	0.0425	0.0409	0.0405	**0.0401**
Emotions	0.2409	0.2450	0.2426	0.2508	0.2450	0.2318
Health	0.0441	0.0456	0.0465	0.0446	**0.0415**	0.0444
Recreation	0.0620	0.0616	0.0630	0.0633	0.0611	**0.0608**
Reference	0.0322	0.0311	0.0345	0.0306	**0.0296**	0.0309
Society	0.0580	0.0577	0.0575	0.0561	0.0559	**0.0545**
Yeast	0.2209	0.2246	**0.2058**	0.2089	0.2080	0.2090
Average	*0.0852*	*0.0859*	*0.0849*	*0.0849*	*0.0857*	**0.0828**

TABLE VIII
COMPARISON BETWEEN MUCO AND OTHER FIVE ALGORITHMS IN TERMS OF F1 (↑)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MUCO
Arts	0.1427	0.1253	0.0880	0.1261	0.1093	**0.1484**
Birds	0.4897	0.4897	0.4897	0.4897	0.0511	**0.4897**
Business	0.6826	0.6856	0.6869	**0.6915**	0.6792	0.6905
Cal500	0.3225	0.3225	0.3240	0.3227	**0.3460**	0.3317
Computer	0.3915	0.4141	0.4210	0.3774	0.4282	**0.4248**
Education	0.1414	0.1797	0.1293	0.2176	0.2070	**0.2448**
Emotions	0.5978	0.5775	0.5689	0.5529	0.5811	**0.6113**
Health	0.3924	0.3269	0.3962	0.3187	**0.4141**	0.3535
Recreation	0.1304	0.1401	0.0930	0.0825	0.1575	**0.1621**
Reference	0.2704	0.3245	0.2782	0.3299	**0.3817**	0.3188
Society	0.2066	0.2034	**0.3499**	0.3009	0.2699	0.3272
Yeast	0.6197	0.6174	**0.6158**	0.6086	0.6212	0.6215
Average	*0.3656*	*0.3672*	*0.3701*	*0.3682*	*0.3539*	**0.3937**

4) for F1, although MUCO is superior to other comparing algorithms on only six datasets, the average performance of the MUCO is better than other comparing algorithms significantly.

From these results shown in Tables III–VIII, we can conclude that MUCO shows better performance compared to the other five state-of-the-art algorithms with different evaluation measures.

To further explore the statistical significance among the six feature selection algorithms, the Friedman test [14] and Bonferroni-Dunn test [12] are employed. Given k comparing algorithms and N datasets, let r_i^j be the rank of the jth algorithm on the ith dataset. $R_i = \frac{1}{N}\sum_{i=1}^{N} r_i^j$ is the average rank of algorithm i among all datasets. Under the null-hypothesis (i.e., all the algorithms are equivalent), the Friedman statistic is distributed according to χ_F^2 with $k-1$ degrees of freedom.

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$
$$\text{where} \quad \chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} R_i^2 - \frac{k(k+1)^2}{4}\right) \quad (22)$$

TABLE IX
SUMMARY OF THE FRIEDMAN STATISTICS F_F ($k = 6, N = 12$) AND THE CRITICAL VALUE ON DIFFERENT EVALUATION MEASURES(k : COMPARING ALGORITHMS; N : DATASETS)

Evaluation Measure	F_F	Critical Value ($\alpha = 0.10$)
AP	4.6301	1.95
RL	4.7277	
CV	3.7486	
OE	4.6832	
HL	2.9017	
F1	2.9472	

where F_F follows a Fisher distribution with $(k-1)$ and $(k-1)(N-1)$ degrees of freedom. Table IX shows the Friedman statistic F_F on different evaluation metrics and the corresponding critical values. According to Table IX, the null hypothesis, which is all algorithms are performing equivalently, is clearly rejected in different evaluation measures at significance level $\alpha = 0.10$. Then, certain posthoc tests such as the Bonferroni–Dunn test can be used to further analyze the relative performance among the comparing algorithms. Here, the difference between the average ranks of MUCO and one baseline is

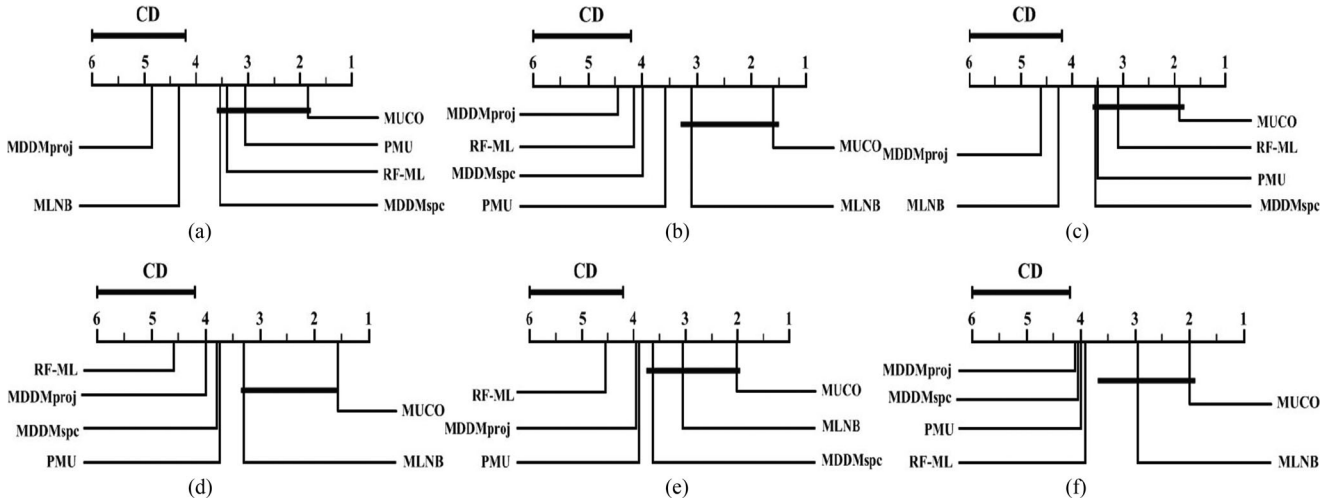

Fig. 7. Comparison of MUCO (control algorithm) against other comparing algorithms with the Bonferroni-Dunn test. (a) RL; (b) OE; (c) CV; (d) AP; (e) HL; (f) F1.

compared with the following *critical difference* (CD):

$$\mathrm{CD}_\alpha = q_\alpha \sqrt{\frac{k(k+1)}{6N}}. \tag{23}$$

Therefore, we have $q_\alpha = 2.326$ at significance level $\alpha = 0.10$, and thus, CD=1.7765 ($k = 6, N = 12$).

To visually show the relative performance of MUCO comparing with other algorithms, Fig. 7 provides the CD diagrams on different evaluation metrics, where the average ranks of each comparing algorithm are plotted along the axis. The lowest (best) ranks on the axis are to the right since we perceive the algorithms on the right side as better. In each subfigure, any comparing algorithm with the average rank within one CD is interconnected with MUCO (the control algorithm). Otherwise, any comparing algorithm whose average rank outside one CD is considered to have significantly different performance with MUCO. Based on the aforementioned results, we can conclude that: 1) for OE, AP, and F1, MUCO performs significantly better than RF-ML, MDDMproj, MDDMspc, and PMU; 2) for HL, MUCO achieves statistically better than MDDMproj, PMU, and RF-ML; and 3) for RL and CV, MUCO is superior to MDDMproj and MLNB.

C. Performance Analysis on MSFS

To verify the effectiveness of MSFS, we compare its performance against static multilabel feature selection methods, i.e., MDDMspc, MDDMproj, RF-ML, PMU, and MLNB. To make our performance comparison authentic and reliable, we use eight benchmark multilabel datasets to effectively simulate streaming features, i.e., features arrive one at a time with a random order, and use the average value of classification performance as the final result after ten random runs. For space considerations, we here selected two datasets to demonstrate the effectiveness of the MSFS. Figs. 8 and 9 show the classification situation with different evaluation measures on Recreation and Reference. There are seven lines in each figures, corresponding to Original, MDDMspc, MDDMproj, RF-ML, PMU, MLNB, and MSFS, respectively. On these lines, Original denotes the final classification performance for all features selected.

TABLE X
NUMBERS OF SELECTED FEATURES WITH MLNB AND MSFS

Dataset	MLNB	MSFS
Computer	344	130
Education	278	173
Enron	482	410
Health	303	166
Recreation	304	181
Reference	406	119
Science	367	207
Society	280	206

From Figs. 8 and 9, we first observe that: 1) for traditional multilabel feature selection, we assume the process of feature selection is conducted on an offline/batch learning manner, and all features of the training set are given *a priori*. The batch manner performs a global search for the best feature at each round, and then, gets a feature rank according to the significance of features; 2) for the MSFS, it assumes that features arrive one at a time, and maintains a best feature subset from the features seen so far by processing each feature upon its arrival. Therefore, we can find that the beginning of the MSFS curve is inferior to traditional multilabel feature selection methods, but it gets better classification performance with a certain number of features. This phenomenon fully shows the difference between MSFS and other baselines and meets the actual streaming configuration.

To demonstrate the effectiveness of the MSFS clearly, we compare the classification performance with five other static multilabel feature selection algorithms, i.e., MDDMspc, MDDMproj, PMU, RF-ML, and MLNB. As MDDMspc, MDDMproj, PMU, and RF-ML get a feature rank list as the result of their feature selection, we select the same number of features determined by MLNB as the final feature subset size, as MLNB gets a feature subset directly. Table X shows the num-

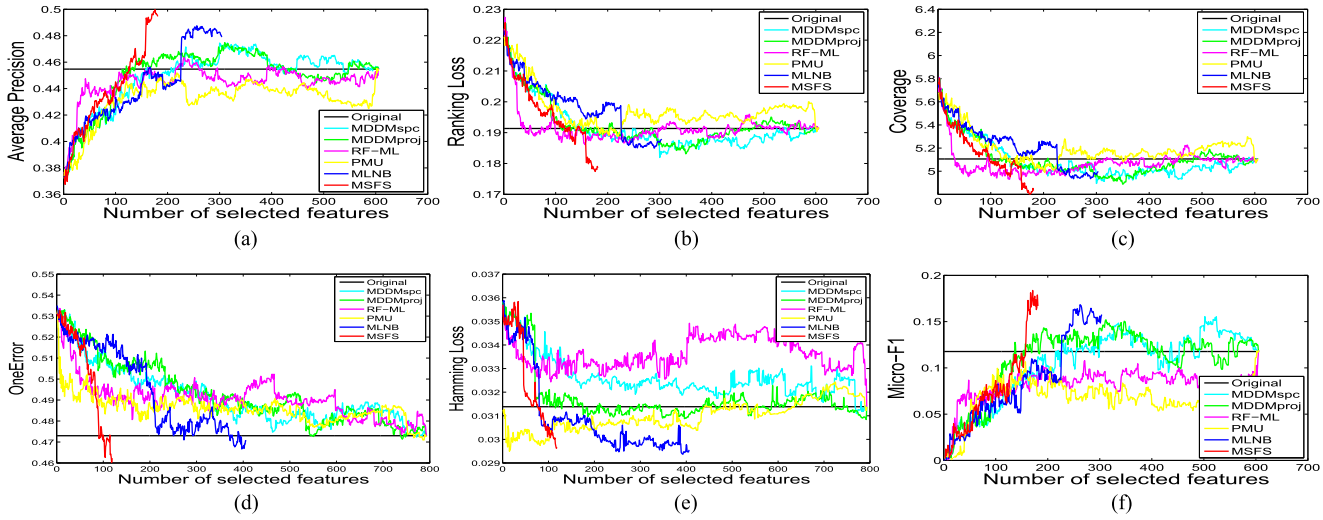

Fig. 8. Classification performance with different evaluation measures on Recreation. (a) AP; (b) RL; (c) CV; (d) OE; (e) HL; (f) F1.

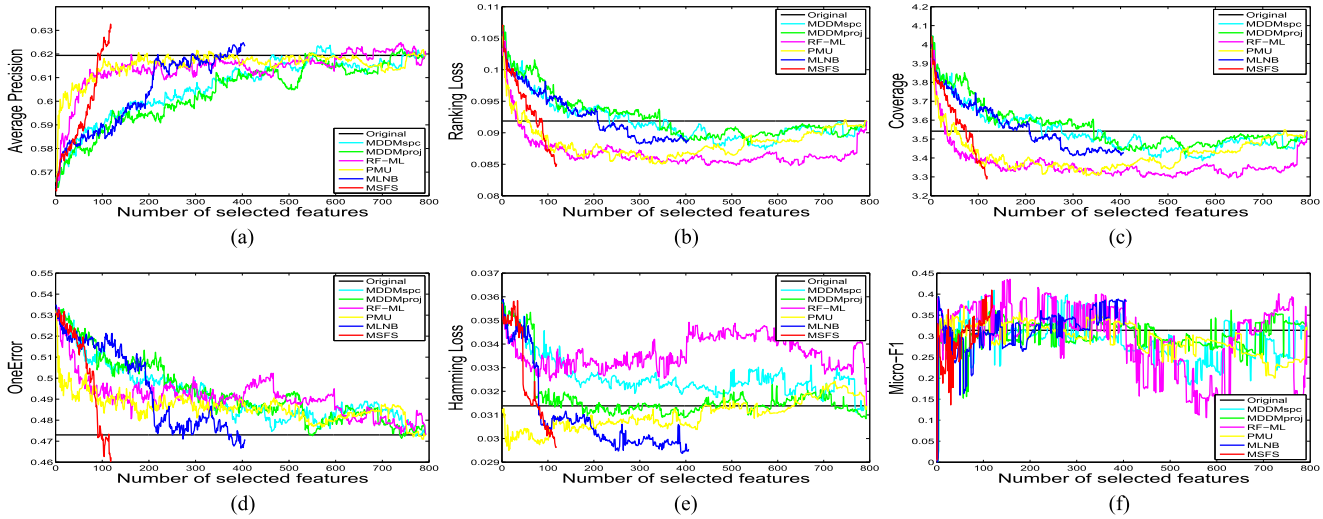

Fig. 9. Classification performance with different evaluation measures on Reference. (a) AP; (b) RL; (c) CV; (d) OE; (e) HL; (f) F1.

TABLE XI
COMPARISON BETWEEN MSFS AND OTHER FIVE ALGORITHMS IN TERMS OF AP (↑)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MSFS
Computer	0.6345	0.6284	0.6285	0.6276	**0.6391**	0.6352
Education	0.5389	0.5425	0.5365	0.5465	0.5478	**0.5595**
Enron	0.6335	0.6179	0.6362	0.6344	0.6242	**0.6474**
Health	0.6654	0.6502	0.6686	0.6802	**0.6880**	0.6848
Recreation	0.4717	0.4703	0.4465	0.4365	0.4790	**0.4946**
Reference	0.6126	0.6106	0.6151	0.6169	0.6234	**0.6308**
Science	0.4547	0.4430	0.4690	0.4416	0.4613	**0.4851**
Society	0.5615	0.5681	0.5900	0.5881	0.5894	**0.5942**
Average	*0.5716*	*0.5664*	*0.5738*	*0.5715*	*0.5815*	**0.5915**

TABLE XII
COMPARISON BETWEEN MSFS AND OTHER FIVE ALGORITHMS IN TERMS OF RL (↓)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MSFS
Computer	0.0916	0.0934	0.0931	0.0941	**0.0910**	0.0931
Education	0.0914	0.0924	0.0939	0.0911	0.0922	**0.0896**
Enron	0.0969	0.0976	0.0928	0.0942	0.0937	**0.0924**
Health	0.0663	0.0698	0.0643	0.0638	0.0641	**0.0606**
Recreation	0.1838	0.1589	0.1917	0.1955	0.1879	**0.1786**
Reference	0.0888	0.0889	0.0856	0.0868	0.0889	**0.0851**
Science	0.1388	0.1417	0.1369	0.1394	0.1364	**0.1295**
Society	0.1500	0.1484	0.1443	0.1442	0.1456	**0.1416**
Average	*0.1135*	*0.1148*	*0.1128*	*0.1136*	*0.1125*	**0.1088**

bers of selected features between MLNB and MSFS. Tables XI–XVI show the classification performance obtained with MLNB, MDDMspc, MDDMproj, PMU, RF-ML, and MSFS, respectively.

For all experimental results in these tables, we can conclude that

1) the average classification performance of the MSFS is superior to comparing algorithms for all evaluation metrics;

TABLE XIII
COMPARISON BETWEEN MSFS AND OTHER FIVE ALGORITHMS IN TERMS OF CV (↓)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MSFS
Computer	4.3987	4.4437	4.4427	4.5013	**4.3740**	4.4700
Education	3.8987	3.9203	3.9920	3.8990	3.9183	**3.8303**
Enron	13.5561	13.5147	13.0466	13.2470	13.1831	**13.0380**
Health	3.5057	3.6217	3.4257	3.4070	3.4163	**3.2800**
Recreation	4.9403	4.9470	5.0860	5.1367	4.9953	**4.8333**
Reference	3.4390	3.4460	3.3270	3.3660	3.4313	**3.3040**
Science	6.9483	7.0840	6.8587	6.9987	6.8367	**6.5600**
Society	5.8423	5.8000	5.6740	5.6603	5.7390	**5.6150**
Average	*5.8161*	*5.8472*	*5.7316*	*5.7770*	*5.7368*	**5.6171**

TABLE XIV
COMPARISON BETWEEN MSFS AND OTHER FIVE ALGORITHMS IN TERMS OF OE (↓)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MSFS
Computer	0.4403	0.4490	0.4483	0.4467	**0.4320**	0.4403
Education	0.6100	0.5973	0.6100	0.5920	0.5827	**0.5707**
Enron	0.2832	0.3178	0.2936	0.2798	0.3161	**0.2729**
Health	0.4270	0.4403	0.4370	0.4080	**0.3947**	0.4093
Recreation	0.6793	0.6827	0.7113	0.7210	0.6643	**0.6490**
Reference	0.4843	0.4887	0.4930	0.4867	0.4703	**0.4623**
Science	0.6823	0.6943	0.6573	0.7010	0.6713	**0.6337**
Society	0.4953	0.4813	0.4593	0.4593	0.4540	**0.4497**
Average	*0.5127*	*0.5189*	*0.5137*	*0.5118*	*0.5076*	**0.4860**

TABLE XV
COMPARISON BETWEEN MSFS AND OTHER FIVE ALGORITHMS IN TERMS OF HL (↓)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MSFS
Computer	0.0406	0.0406	0.0421	0.0413	0.0401	**0.0394**
Education	0.0426	0.0422	0.0425	0.0409	**0.0405**	0.0407
Enron	0.0522	0.0527	0.0524	0.0518	0.0525	**0.0509**
Health	0.0441	0.0456	0.0465	0.0446	**0.0415**	0.0420
Recreation	0.0620	0.0616	0.0630	0.0633	0.0611	**0.0607**
Reference	0.0322	0.0311	0.0345	0.0306	0.0296	**0.0296**
Science	0.0347	0.0348	**0.0342**	0.0352	0.0346	0.0343
Society	0.0580	0.0577	0.0575	0.0561	**0.0559**	0.0560
Average	*0.0458*	*0.0458*	*0.0466*	*0.0455*	*0.0455*	**0.0422**

TABLE XVI
COMPARISON BETWEEN MSFS AND OTHER FIVE ALGORITHMS IN TERMS OF F1 (↑)

Dataset	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	MSFS
Computer	0.3915	0.4141	0.4210	0.3774	**0.4282**	0.3847
Education	0.1414	0.1797	0.1293	0.2176	0.2070	**0.2414**
Enron	0.4745	0.4138	0.4777	0.4881	0.4597	**0.4829**
Health	0.3924	0.3269	0.3962	0.3187	0.4141	**0.4144**
Recreation	0.1304	0.1401	0.0930	0.0825	0.1575	**0.1669**
Reference	0.2704	0.3245	0.2782	0.3299	**0.3817**	0.3661
Science	0.0915	0.0736	0.1282	0.0658	0.1128	**0.1225**
Society	0.2066	0.2034	**0.3499**	0.3009	0.2699	0.2749
Average	*0.2623*	*0.2595*	*0.2842*	*0.2726*	*0.3039*	**0.3067**

TABLE XVII
SUMMARY OF THE FRIEDMAN STATISTICS F_F ($k = 6, N = 8$) AND THE CRITICAL VALUE ON DIFFERENT EVALUATION MEASURES (k : COMPARING ALGORITHMS; N : DATASETS)

Evaluation Measure	F_F	Critical Value ($\alpha = 0.10$)
AP	10.0433	2.00
RL	3.6089	
CV	4.4285	
OE	8.6956	
HL	7.2029	
F1	2.7756	

2) for the label ranking evaluation performance, MSFS gets better performance against all comparing algorithms on at least six datasets. Note that the classification performance of the MSFS is extremely close with the best value on other two datasets for all evaluation metrics;
3) for the label set prediction accuracy, MSFS is superior to other comparing algorithms on the average performance, and obtains the best value on at least four datasets;
4) the number of selected features with MSFS is less than MLNB and other comparing algorithms, but MSFS achieves superior or at least comparable performance against all comparing algorithms.

These results indicate that MSFS performs better than all baselines when facing streaming features.

To provide performance analysis among the comparing algorithms in a systematical way, we also employ the Friedman test [14] and Bonferroni–Dunn test [12] to perform statistical analysis. For Bonferroni–Dunn test, we have $q_\alpha = 2.326$ at significance level $\alpha = 0.10$, and thus, CD $= 2.1757$ ($k = 6, N = 8$), as shown in Table XVII. Accordingly, the performance between the MSFS and a comparing algorithm is deemed to be significantly different if their average ranks over all datasets differ by at least one CD.

To visually illustrate the relative performance of the MSFS and other comparing algorithms, Fig. 10 shows the CD diagrams on different evaluation measures. From Fig. 10, we can conclude that

1) for AP, OE, HL, and RL, MSFS significantly outperforms MDDMspc, MDDMproj, PMU, RF-ML, and MLNB, and at least obtains comparable performance against MLNB;
2) for CV and F1, MSFS achieves statistically better than MDDMspc, MDDMproj, and PMU, and obtains comparable performance against RF-ML and MLNB;
3) across all evaluation metrics, MSFS carries comparable performance against MLNB, however, the number of selected features of the MSFS is far fewer than MLNB, as shown in Table X.

To summarize, MSFS provides highly competitive performance against comparing algorithms when facing streaming features.

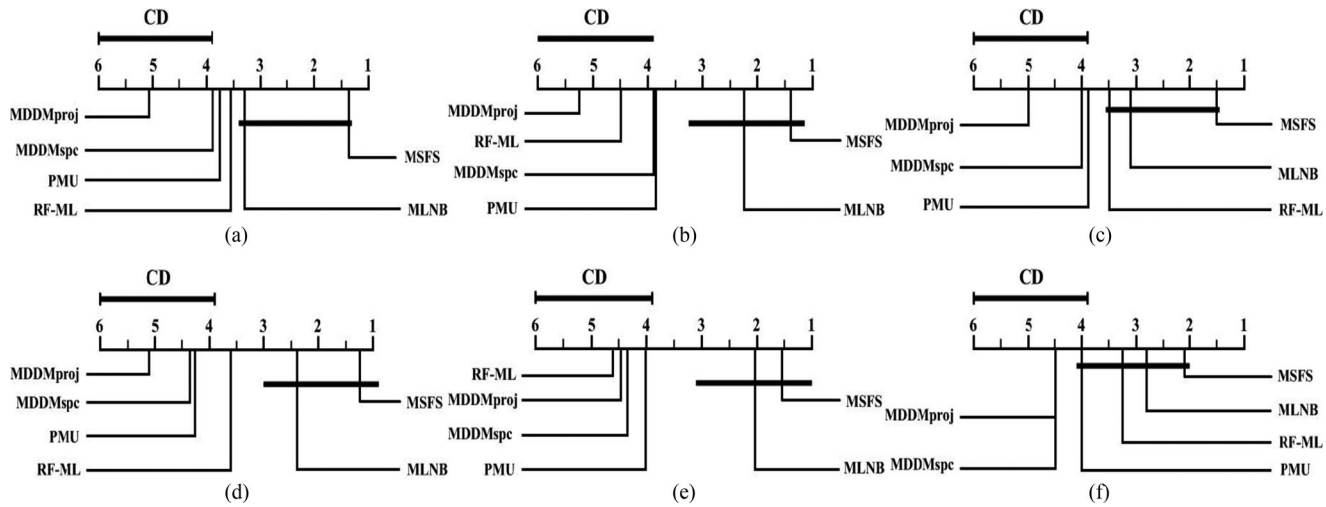

Fig. 10. Comparison of the MSFS (control algorithm) against other comparing algorithms with the Bonferroni–Dunn test. (a) RL; (b) OE; (c) CV; (d) AP; (e) HL; (f) F1.

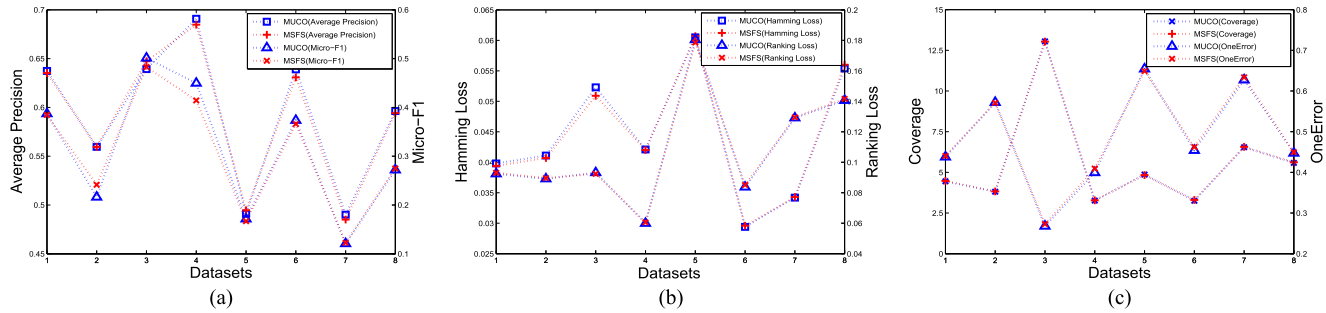

Fig. 11. Comparisons between MUCO and MSFS (the labels of the x-axis from 1 to 8 denote the datasets: 1: Computer; 2: Education; 3: Enron; 4: Health; 5: Recreation; 6: Reference; 7: Science; 8: Society). (a) AP and F1. (b) HL and RL. (c) CV and OE.

D. Comparisons Between MUCO and MSFS

A comparison between MUCO and MSFS in Fig. 11 demonstrates that MSFS is a little inferior to MUCO, but selects less features than MUCO, as MUCO obtains a feature rank list ultimately. Both MUCO and MSFS perform multiple statistical comparisons to assess whether a feature is relevant or redundant. In the relevance and redundancy analysis phrase, MUCO needs to evaluate each feature within the whole feature space. MSFS, on the other hand, can significantly reduce the total number of comparisons, because it first examines the relevance of a new feature, and then, checks the redundancy between the new feature and only one feature of selected features.

Based on the comparative analysis between MUCO and MSFS, we can use MSFS when features are no longer static but flow in one by one, and each new feature needs to be processed upon its arrival. Different from MSFS, MUCO needs to wait a long time for all features to become available, and then, carries out multilabel feature selection.

E. Runtime Analysis

To show the computational efficiency of our proposed algorithms, in this section, we give a comparison on efficiency among PMU, MUCO, and MSFS. Because all of these three multilabel feature selection algorithms are based on the information theory. Moreover, for illustrating the results impartially and clearly, we select six datasets that all exist in Tables I and II, i.e., Computer, Education, Health, Recreation, Reference, and Society. In addition, the hardware platform for our experiments is a PC equipped with 32-G main memory and 3.1-GHZ CPU. The software is Windows 7 and MATLAB (Version 2012a). The results in Table XVIII show that: 1) PMU is faster than MUCO, because the runtime of MUCO is significantly influenced by the computation of fuzzy mutual information and the Max-Relevance and Min-Redundancy strategy, but PMU

TABLE XVIII
RUNTIME ANALYSIS (S) OF PMU, MUCO, AND MSFS

Dataset	PMU	MUCO	MSFS
Computer	9268	123633	5451
Education	6246	72403	4440
Health	25043	92834	4984
Recreation	51342	92270	4849
Reference	24162	174233	6294
Society	9925	104166	5166

maximizes the dependence between the selected features and labels only considers 3-D interactions among features and labels; and 2) MSFS is much faster than MUCO on all datasets, because MSFS selects an effective feature subset by online analysis, and does not need to obtain the whole feature space in advance.

V. Conclusion

When the knowledge of the full feature space is either known or unknown in advance, in this paper, we have presented two new algorithms based on fuzzy mutual information for multilabel feature selection: MUCO and MSFS, respectively. MUCO addresses label correlation and feature selection simultaneously, and MSFS solves label correlation, streaming features, and feature selection in one shot. Compared to the five state-of-the-art methods, MDDMspc, MDDMproj, PMU, RF-ML, and MLNB, the presented algorithms MUCO and MSFS have shown that they can measure the quality of features effectively. In the experiments, our study has shown that:

1) for a known feature space, MUCO can obtain a better feature list via the strategy of Max-Relevance and Min-Redundancy;
2) for an unknown feature space, MSFS can select a small number of features to train a much stronger model;
3) the experiments have demonstrated that the prediction accuracy of the proposed algorithms is mostly higher than, or at least as good as, other methods.

In some real-world applications, such as drug repositioning and image analysis, features may arrive by groups. Therefore, in future work, we will study online multilabel group feature selection.

Acknowledgment

The authors would like to thank the anonymous reviewers and the editor for their constructive and valuable comments.

References

[1] A. Alalga, K. Benabdeslem, and N. Taleb, "Soft-constrained Laplacian score for semi-supervised multi-label feature selection," *Knowl. Inf. Syst.*, vol. 47, pp. 75–98, 2016.
[2] D. Bell and H. Wang, "A formalism for relevance and its application in feature subset selection," *Mach. Learn.*, vol. 41, pp. 175–195, 2000.
[3] M. Boutell, J. Luo, X. Shen, and C. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, pp. 1757–1771, 2004.
[4] F. Charte, D. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "R ultimate multilabel dataset repository," in *Proc. 11th Int. Conf. Hybrid Artif. Intell. Syst.*, vol. 9648, 2016, pp. 1–13.
[5] F. Charte, D. Charte. "Working with Multilabel Datasets in R: The mldr package," *R Journal*, vol. 7, no.2, pp. 149–162, 2015.
[6] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, pp. 3–16, 2015.
[7] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowl.-Based Syst.*, vol. 89, pp. 385–397, 2015.
[8] H. Chen, T. Li, C. Luo, S.-J. Horng and G. Wang, "A decision-theoretic rough set approach for dynamic data mining," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 6, pp. 1958–1970, Dec. 2015.
[9] D. Chen and Y. Yang, "Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 5, pp. 1325–1334, Oct. 2014.
[10] W. Ding *et al.*, "Sub-kilometer crater discovery with boosting and transfer learning," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 4, pp. 1–22, 2011.
[11] G. Drzadzewski and F. Tompa, "Partial materialization for online analytical processing over multi-tagged document collections," *Knowl. Inf. Syst.*, vol. 47, pp. 697–732, 2016.
[12] O. Dunn, "Multiple comparisons among means," *J. Amer. Stat. Assoc.*, vol. 56, pp. 52–64, 1961.
[13] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. 14th Int. Conf.Adv. Neural Inf. Process. Syst.*, Cambridge, MA, USA, 2002, pp. 681–687.
[14] M. Friedman, "A comparison of alternative tests of significance for the problem of m ranking," *Ann. Math. Stat.*, vol. 11, pp. 86–92, 1940.
[15] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, NY, USA: Academic, 1990.
[16] A. Gani, A. Siddiqa, S. Shamshirband, and F. Hanum, "A survey on indexing techniques for big data: Taxonomy and performance evaluation," *Knowl. Inf. Syst.*, vol. 46, pp. 241–284, 2016.
[17] W. Gao and Z. Zhou, "On the consistency of multi-label learning," *Artif. Intell.*, vol. 199, pp. 22–44, 2013.
[18] O. Gharroudi, H. Elghazel, and A. Aussem, "A comparison of multi-label feature selection methods using the random forest paradigm," in, *Advances in Artificial Intelligence*. Cham, Switzerland: Springer, pp. 95–106, 2014.
[19] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. Cham, Switzerland: Springer, 2015.
[20] E. Gibaja and S. Ventura, "Multi-label learning: A review of the state of the art and ongoing research," *Wiley Interdiscip. Rev. Data Mining Knowl. Discovery*, vol. 4, no. 6, pp. 411–444, 2014.
[21] Q. Gu, Z. Li, and J. Han, "Correlated multi-label feature selection," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 1087–1096.
[22] B. Hariharan, S. Vishwanathan, and M. Varma, "Efficient max-margin multi-label classification with applications to zero-shot learning," *Mach. Learn.*, vol. 88, no. 1–2, pp. 127–155, 2012.
[23] H. Hotelling, "Relations between two sets of variables," *Biometrika*, vol. 28, pp. 312–377, 1936.
[24] Q. Hu, D. Yu, Z. Xie, and J. Liu, "Fuzzy probabilistic approximation spaces and their information measures," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 2, pp. 191–201, Apr. 2006.
[25] Q. Hu *et al.*, "Feature selection for monotonic classification," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 1, pp. 69–81, Feb. 2012.
[26] S. Huang and Z. Zhou, "Multi-label learning by exploiting label correlations locally," in *Proc. 26th AAAI Conf. Artif. Intell.*, Toronto, Canada, 2012, pp. 949–955.
[27] M. Javidi and S. Eskandari, "Streamwise feature selection: A rough set method," *Int. J. Mach. Learn. Cybern.*, vol. 1, pp. 1–10, 2016.
[28] V. Kumar and S. Minz, "Multi-view ensemble learning: An optimal feature set partitioning for high-dimensional data classification," *Knowl. Inf. Syst.*, vol. 49, pp. 1–59, 2016.
[29] J. Lee and D. Kim, "Feature selection for multi-label classification using multivariate mutual information," *Pattern Recognit. Lett.*, vol. 34, pp. 349–357, 2013.
[30] J. Lee and D. Kim, "Fast multi-label feature selection based on information-theoretic feature ranking," *Pattern Recognit.*, vol. 48, no. 9, pp. 2761–2771, 2015.
[31] P. Li, H. Li, and M. Wu, "Multi-label ensemble based on variable pairwise constraint projection," *Inf. Sci.*, vol. 222, pp. 269–281, 2013.
[32] Y. Lin, Q. Hu, J. Liu, J. Chen, and J. Duan, "Multi-label feature selection based on neighborhood mutual information," *Appl. Soft Comput.*, vol. 38, pp. 244–256, 2016.
[33] Y. Lin, Q. Hu, J. Liu, and J. Duan, "Multi-label feature selection based on max-dependency and min-redundancy," *Neurocomputing*, vol. 168, pp. 92–103, 2015.
[34] W. Liu and T. Wang, "Online active multi-field learning for efficient email spam filtering," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 117–136, 2012.
[35] H. Liu, S. Zhang, and X. Wu, "MLSLR: Multilabel learning via sparse logistic regression," *Inf. Sci.*, vol. 281, pp. 310–320, 2014.
[36] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[37] S. Perkins and J. Theiler, "Online feature selection using grafting," in *Proc. 20th Int. Conf. Mach. Learn.*, Washington, DC, USA, 2003, pp. 592–599.
[38] Y. Qian, F. Li, J. Liang, B. Liu, and C. Dang, "Space structure and clustering of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2047–2059, Oct. 2015, doi: 10.1109/TNNLS.2015.2451151.
[39] J. Read, "A pruned problem transformation method for multi-label classification," in *Proc. New Zealand Comput. Sci. Res. Student Conf.*, 2008, pp. 143–150.
[40] R. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Mach. Learn.*, vol. 39, no. 2, pp. 135–168, 2000.
[41] N. Spolaõr, E. Cherman, and M. Monard, "Using ReliefF for multi-label feature selection," in *Proc. Conf. Latinoamericana de Informatica*, 2011, pp. 960–975.
[42] M. A. Tahir, J. Kittler, and F. Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification," *Pattern Recognit.*, vol. 45, no. 10, pp. 3738–3750, 2012.
[43] K. Trohidis, G. Tsoumakas, G. Kalliris, I. Katakis, and I. Vlahavas, "Multi-label classification of music into emotions," in *Proc. 9th Int. Soc. Music Inf. Retrieval.*, Philadelphia, PA, USA, 2008, pp. 325–330.
[44] G. Tsoumakas, E. Spyromitros-Xiousfis, and I. Vilcek, "Mulan: A java library for multi-label learning," *J. Mach. Learn. Res.*, vol. 12, pp. 2411–2414, 2011.
[45] S. Wang, J. Wang, Z. Wang, and Q. Ji, "Enhancing multi-label classification by modeling dependencies among labels," *Pattern Recognit.*, vol. 47, pp. 3405–3413, 2014.
[46] X. Wu, X. Zhu, G. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.
[47] X. Wu, K. Yu, H. Wang, and W. Ding, "Online streaming feature selection," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 1159–1166.
[48] M. Xu, Y. Li, and Z. Zhou, "Multi-label learning with PRO loss," in *Proc. 27th AAAI Conf. Artif. Intell.*, Bellevue, WA, USA, 2013, pp. 998–1004.
[49] D. Yu, S. An, and Q. Hu, "Fuzzy mutual information based min-redundancy and max-relevance heterogeneous feature selection," *Int. J. Comput. Intell. Syst.*, vol. 4, pp. 619–633, 2011.
[50] J. Yu and W. Xu, "Incremental knowledge discovering in interval-valued decision information system with the dynamic data," *Int. J. Mach. Learn. Cybern.*, vol. 8, pp. 849–864, 2017.
[51] K. Yu, X. Wu, W. Ding, and J. Pei, "Towards scalable and accurate online feature selection for big data," in *Proc. IEEE 14th Int. Conf. Data Mining*, Shenzhen, China, 2014, pp. 661–669.
[52] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, 2005, pp. 258–265.
[53] Y. Yu, W. Pedrycz, and D. Miao, "Multi-label classification by exploiting label correlations," *Expert Syst. Appl.*, vol. 41, pp. 2989–3004, 2014.
[54] L. Zhang, Q. Hu, J. Duan, and X. Wang, "Multi-label feature selection with fuzzy rough sets," in *Rough Sets and Knowledge Technology*. Cham, Switzerland: Springer, 2014, pp. 121–128.
[55] M. Zhang and Z. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Inf. Sci.*, vol. 40, pp. 2038–2048, 2007.
[56] M. Zhang, J. Peria, and V. Robles, "Feature selection for multi-label naive Bayes classificaiton," *Inf. Sci.*, vol. 179, pp. 3218–3229, 2009.
[57] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[58] M. Zhang, and L. Wu, "LIFT: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.
[59] Y. Zhang and Z. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Trans. Knowl. Discovery Data*, vol. 4, pp. 1–21, 2010.
[60] J. Zhang, Z. Zhao, X. Hu, Y. Cheung, M. Wang, and X. Wu, "Online group feature selection," in *Proc. 23rd. Int. Joint Conf. Artif. Intell.*, Beijing, China, 2013, pp. 1757–1763.
[61] J. Zhou, D. Foster, R. Stine, and L. Ungar, "Streamwise feature selection using alpha-investing," in *Proc. 11th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Chicago, IL, USA, 2005, pp. 384–393.
[62] T. Zhu, G. Li, W. Zhou, P. Xiong, and C. Yuan, "Privacy-preserving topic model for tagging recommender systems," *Knowl. Inf. Syst.*, vol. 46, pp. 33–58, 2016.

Yaojin Lin received the Ph.D. degree in computers with applications from the School of Computer and Information, Hefei University of Technology, Hefei, China, in 2014.

He is currently an Associate Professor with Minnan Normal University, Zhangzhou, China, and a Postdoctoral Fellow with Tianjin University, Tianjin, China. His research interests include data mining and granular computing. He has published more than 30 papers in many journals, such as *Neurocomputing*, *Decision Support Systems*, *Information Sciences*, and *Applied Intelligence*.

Qinghua Hu (SM'13) received the B.S., M.S., and Ph.D. degrees in computers with applications from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively.

He is currently a Professor with Tianjin University, Tianjin, China. His research interests include machine learning, data mining, and granular computing. He has published more than 90 journal and conference papers in the areas of data mining and machine learning.

Jinghua Liu is currently working toward the Master degree with the School of Computer Science, Minnan Normal University, Zhangzhou, China.

Her research interests include data mining and granular computing

Jinjin Li received the M.S. degree in mathematics from Guangxi University, Guangxi, China, in 1988, and the Ph.D. degree in fundamental mathematics-from the School of Mathematics and System Sciences, Shandong University, Shandong, China, in 2000.

He is currently a Professor with the School of Mathematics and Statistics, Minnan Normal University, Zhangzhou, China. He has published more than 70 articles in international journals and book chapters. His research interests include the area of data mining, rough sets, and topologies.

Xindong Wu (SM'95–F'11) received the Ph.D. degree in artificial intelligence from the University of Edinburgh, Edinburgh, U.K.

He is Alfred and Helen Lamson Endowed Professor in Computer Science in the School of Computing and Informatics at the University of Louisiana at Lafayette, LA, USA. His research interests include data mining, knowledge-based systems, and Web information exploration. He has published more than 200 refereed papers as well as 25 books and conference proceedings in these areas. His research has been supported by the U.S. National Science Foundation, the U.S. Department of Defense, the National Natural Science Foundation of China, and the Chinese Academy of Sciences, as well as industrial companies including Microsoft Research, U.S. West Advanced Technologies and Impact Solutions.

Dr. Wu is the Founder and current Steering Committee Chair of the IEEE International Conference on Data Mining, the Founder and current Editor-in-Chief of *Knowledge and Information Systems* (Springer), the Founding Chair (2002–2006) of the IEEE Computer Society Technical Committee on Intelligent Informatics, and a Series Editor of the Springer Book Series on *Advanced Information and Knowledge Processing*. He was the Editor-in-Chief of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. He served as the Program Committee Chair/Cochair for the 2003 IEEE International Conference on Data Mining, the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, and the 19th ACM Conference on Information and Knowledge Management.