Pattern Recognition 66 (2017) 342–352

Contents lists available at ScienceDirect

Pattern Recognition

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

SCLS: Multi-label feature selection based on scalable criterion for large label set

Jaesung Lee, Dae-Won Kim*

School of Computer Science and Engineering, Chung-Ang University, 221, Heukseok-Dong, Dongjak-Gu, Seoul 156-756, Republic of Korea

ARTICLE INFO

Keywords:
Machine learning
Multi-label learning
Multi-label feature selection
Relevance evaluation
Conditional relevance

ABSTRACT

Multi-label feature selection involves the selection of relevant features from multi-labeled datasets, resulting in a potential improvement of multi-label learning accuracy. In conventional multi-label feature selection methods, the final feature subset is obtained by identifying the features of high relevance with low redundancy. Thus, accurate score evaluation is a key factor for obtaining an effective feature subset. However, conventional methods suffer from inaccurate conditional relevance evaluation when a large number of labels are involved. As a result, irrelevant features can be a member of the final feature subset, leading to low multi-label learning accuracy. In this paper, we propose a new multi-label feature selection method. Using a scalable relevance evaluation process that evaluates conditional relevance more accurately, the proposed method significantly improves multi-label learning accuracy compared with conventional multi-label feature selection methods.

1. Introduction

Multi-label classification is part of a base technique for recent applications, such as sentiment analysis of user texts [25,28] or tag classification of music clips [24,35,39,41], because texts and music clips can be associated with multiple concurrent labels [20,43]. In practice, applications can incur a series of labels for encoding the target concepts to be learned, especially when the target consists of multiple sub-concepts, such as humor or admiration [1,8,12]. Let $W \subset \mathbb{R}^d$ denote a set of patterns constructed from a set of features F. Then each pattern $w_i \in W$ where $1 \le i \le |W|$ is assigned to a certain label subset $\lambda_i \subseteq L$, where $L = \{l_1, \ldots, l_{|L|}\}$ and is a finite set of labels. Because there can be hidden relationships among these tags or labels that would improve multi-label learning accuracy, better performance can be achieved by exploiting useful relationships [36,44]. For this reason, multi-label feature selection can contribute to the improvement of learning accuracy by highlighting such relationships based on important features [17,33,42], and hence it is considered an important preprocessing step [16,18,19].

Given input data with an original feature set F and label set L, the goal of multi-label feature selection is to identify a feature subset $S \subset F$ with $n \ll |F|$ features that have the largest relevance on multiple labels [16,19]. Because S should support multiple labels simultaneously using only n features, the selection of a compact feature subset becomes an important task when L involves many labels [18,21,22,23]. To ensure the largest relevance on L with n features, each feature in S should carry individual information on labels [19]. If two features carry the same information, it becomes unnecessary to select one of them to compose S because this feature will not carry any additional discriminating power under the selection of the other feature. Thus, relevance evaluation that considers dependency among the selected features is important for identifying an effective feature subset.

Under the incremental selection for efficiently finding a near-optimal solution, the selection of the i-th feature from the set $\{F - S_{i-1}\}$, where S_{i-1} is a feature subset with $i-1$ features, is performed by identifying the f_i that maximizes the value of following the relevance criterion [16,19,21,23].

$$\max_{f_i \in \{F - S_{i-1}\}} [Rel(f_i) - Red(f_i)] \tag{1}$$

where $Rel(f_i)$ and $Red(f_i)$ denote the dependency of f_i to L and the dependency between f_i and the already selected features of S_{i-1}, respectively. Thus, the task can be solved by scoring each feature based on Eq. (1), and then including the top-ranked feature at each iteration [13,17,32]. In the multi-label feature selection method that employs Eq. (1), the algorithm attempts to avoid the selection of features carrying the same information that is given by already selected features based on $Red(f_i)$.

In previous studies, $Rel(f_i)$ is calculated as a large value because it is computed by adding the dependency values between f_i and each label. On the other hand, when the number of involved labels is large, $Red(f_i)$ implies too small values compared with $Rel(f_i)$ [21,22], or incurs

* Corresponding author.
E-mail address: dwkim@cau.ac.kr (D.-W. Kim).

http://dx.doi.org/10.1016/j.patcog.2017.01.014
Received 12 August 2016; Received in revised form 13 December 2016; Accepted 9 January 2017
Available online 10 January 2017
0031-3203/

erroneous calculation because of repetitive dependency computations along with each label [16,23]. As a result, irrelevant features can be included in the final feature subset because of inaccurate relevance evaluation. In this paper, we propose an effective multi-label feature selection method based on a new $Red(f_i)$ function that considers $Rel(f_i)$ within its calculation while avoiding erroneous dependency calculations, resulting in the improvement of multi-label learning accuracy.

2. Related work

In multi-label feature selection studies, one of the major trends includes the application of a feature selection method for single-label problems after transforming label sets into a single label [29,33]. In addition to the merits from the immediate use of conventional methods and their side effects [34], an algorithm adaptation strategy that directly manages multi-label problems has also been considered [36]. In this approach, a feature subset is obtained by the optimization of a certain criterion, such as a joint learning criterion that involves simultaneous feature selection and multi-label learning [10,27], $l_{2,1}-$norm function optimization [26], label ranking error [9], Hilbert-Schmidt independence criterion [14], F-statistics [13], and label-specific feature selection [43]. In this paper, we focus on a mutual information-based multi-label feature selection method because its theoretical background and advantage have been thoroughly discussed in previous studies [6,16,17,21–23].

When mutual information is employed in its original form for evaluating feature relevance, the algorithm inevitably faces the problem of high-dimensional joint probability estimation caused by multiple labels in L [16,21]. Because the process often becomes impractical given the insufficient patterns and characteristics of particular labels [1,3], researchers have attempted to circumvent this difficulty by focusing on dependency among variable subsets [19,33], resulting in variations that provide each with a unique advantage against the characteristics of datasets and evaluation measures [18].

In the work of [16], the authors demonstrated that mutual information can be decomposed into the sum of dependencies among all possible variable subsets across S and L. To circumvent the intractable calculations, the dependency between features and the label set is approximated by considering each label and label pairs. In addition, the dependency among features is determined by adding the dependency of all combinations composed of two features and one label. Finally, the relevance of the feature is calculated by subtracting the dependency among the features from that to labels. A similar approach was also employed in the pieces of works of [19,23]. Thus, relevance evaluation is commonly based on calculations that involve repetitive dependency estimates for too many variable subsets along with given labels. As a result, the relevance evaluation process is unscalable to the number of labels because possible errors caused from the dependency estimation for each variable subset will be cumulated to the final relevance score.

If two features are mutually independent, these two features certainly have a different nature in terms of dependency on L. Based on this property, a criterion for relevance evaluation can be derived without incurring repetitive dependency calculations [21]. The same score function was employed for a quadratically programmed objective function for considering the global perspective of a selected feature subset [22]. However, these two methods commonly suffer from an incompact feature subset because feature dependency is determined regardless of the amount of dependency on labels, resulting in underestimation of feature dependency compared with the dependency on labels when a large number of labels are involved.

3. Proposed method

3.1. Limitations of previous studies

As Eq. (1) shows, the characteristics of the selected feature subset is strongly influenced by a relevance evaluation based on $Rel(f_i)$ and $Red(f_i)$. For example, if $Rel(f_i)$ is evaluated as too large compared with $Red(f_i)$, the influence of $Red(f_i)$ on the relevance evaluation eventually becomes small. In this case, the selected feature subset can be composed of features that are dependent on each other, resulting in low discriminating power within a fixed number of features. This undesirable situation can occur from conventional multi-label feature selection methods when the number of labels is large [21,22] because $Rel(f_i)$ increases as the number of labels grows, whereas $Red(f_i)$ is solely determined by the dependency on features in S_{i-1} [21,22].

To show this aspect more clearly, we conduct a preliminary analysis in this section. For demonstration purposes, we select a conventional multi-label feature selection method from the perspective of simplicity. In the work of [21,22], the i-th feature f_i is selected if it maximizes

$$\max_{f_i\in\{F-S_{i-1}\}}\left[\sum_{l\in L} M(f_i;l)-\sum_{f\in S_{i-1}} M(f_i;f)\right] \tag{2}$$

where $M(x;y)=H(x)-H(x,y)+H(y)$ is the mutual information between variables x and y, and $H(x)=-\sum P(x)\log P(x)$ is the joint entropy with their probability functions $P(x)$, $P(y)$, and $P(x,y)$. Eq. (2) indicates that $Rel(f_i)=\sum_{l\in L} M(f_i;l)$ and $Red(f_i)=\sum_{f\in S_{i-1}} M(f_i;f)$ are respectively implemented as the sum of mutual information terms: (1) between f_i and all the labels, and (2) between f_i and the already selected features in S_{i-1}. Thus, the number of mutual information terms considered in $Rel(f_i)$ is the same as the number of labels in L. In contrast, the number of the terms in $Red(f_i)$ is $|S_{i-1}|$. For example, let us consider $i=2$ where the algorithm selects the second feature. Because the number of mutual information terms in $Rel(f_i)$ is fixed to $|L|$, whereas that in $Red(f_i)$ is only one, the influence of $Red(f_i)$ on the relevance evaluation is small. As a result, a feature that is dependent on the already selected feature can be selected to compose S_2. The example shows that the feature subset with a small number of features can be composed of dependent features that are unhelpful to the improvement of multi-label learning accuracy. For multi-label feature selection, this is a serious problem because the number of features to be selected in n is typically set to a small number. Moreover, this example also indicates that the conventional method requires more features in order to attain multi-label learning accuracy to some extent because the dependent features could be members of the final feature subset, and non-influential on the improvement of multi-label learning accuracy. The reason for this result is that $Red(f_i)$ is determined regardless of $Rel(f_i)$, hence this problem can be solved by adjusting the influence of $Red(f_i)$ against $Rel(f_i)$, or vice versa. Although the $Red(f_i)$ term was calculated differently in the works of [16,19,23], they commonly suffer from too many complex relationships among the features and label combinations, resulting in inaccurate relevance evaluation caused by cumulative error from the probability estimation of dependency calculations.

In this paper, we propose a multi-label feature selection method based on a Scalable Criterion for Large Label Set (SCLS) aimed at identifying an effective feature subset for improving multi-label learning accuracy. The difference between SCLS and previous studies can be summarized as follows:

- SCLS is designed to use a simpler dependency calculation process. For example, in previous studies [16,19], $Red(f_i)$ is calculated as

$$Red(f_i)=\sum_{f\in S_{i-1}}\sum_{l\in L}\left(H(f_i)+H(f)+H(l)-H(f_i,f)-H(f_i,l)\right.$$
$$\left.-H(f,l)+H(f_i,f,l)\right) \tag{3}$$

which involves the computation of feature dependency with regard to each label. In contrast, SCLS avoids repetitive calculations along with each label based on the dependency between f_i and $f \in S_{i-1}$.
- SCLS is designed to consider $Rel(f_i)$ in the calculation of $Red(f_i)$, thus making it scalable to $Rel(f_i)$. As a result, the problem of non-influential $Red(f_i)$ in previous studies [21,22] can be solved by SCLS.
- To achieve a scalable relevance evaluation, SCLS employs an effective approximation for the dependency calculations [15] that is still not considered in the multi-label feature selection problem.

3.2. Modification selection between $Rel(f_i)$ and $Red(f_i)$

Suppose that the algorithm selects the first feature f_1 from F for the first step under the incremental selection strategy. In this step, the calculation of feature dependency on previously selected features from relevance evaluation is unnecessary because $S = \{\varnothing\}$. Thus, the relevance of f_1 is evaluated as [17].

$$M(f_1; L) = H(f_1) - H(f_1, L) + H(L) \tag{4}$$

To circumvent the risk of inaccurate probability estimation caused by high dimensionality of the label set L, Eq. (4) can be rewritten as [16,17].

$$M(f_1; L) = \sum_{k=2}^{|L|+1} (-1)^k V_k(f_1 \times L'_{k-1}) \tag{5}$$

where $\times$ is a Cartesian product between two sets, and $V_k(\cdot)$ is the sum of a k-degree interaction defined as [19].

$$V_k(L') = \sum_{X \in L'_k} I(X) \tag{6}$$

where L' is a power set of L without $\{\varnothing\}$, X is a possible element from $L'_k = \{e | e \in L', |e| = k\}$, and $I(X)$ is the interaction information defined as [4].

$$I(X) = -\sum_{Y \in X'} (-1)^{|Y|} H(Y) \tag{7}$$

Eq. (5) indicates that the interaction information between f_1 and all possible label subsets drawn from L should be calculated in order to obtain the dependency value for f_1. To remedy the intractable computational cost with regard to $|L|$, Eq. (5) can be approximated by setting a parameter that adjusts the maximum allowed cardinality of the label subsets as follows [17,19]:

$$\widetilde{M}_b(f_1; L) = \sum_{k=2}^{b+1} (-1)^k V_k(f_1 \times L'_{k-1}) \tag{8}$$

where $b \geq 1$. As a result, the simplest approximation of Eq. (5) can be obtained by setting b=1 as follows:

$$\widetilde{M}_1(f_1; L) = \sum_{k=2}^{2} (-1)^k V_k(f_1 \times L'_{k-1}) = \sum_{l \in L} M(f_1; l) \tag{9}$$

The relevance of a feature when $S = \{\varnothing\}$, $Rel(f_1)$ in Eq. (1) can be calculated by Eq. (9). The equation takes the same form for relevance evaluation when $S = \{\varnothing\}$ in most mutual information-based multi-label feature selection methods [17,19,21–23]. Moreover, Eq. (9) indicates that $Rel(f_1)$ can be a large value when $|L|$ is large because $0 \leq M(f_1; l)$, and the dependency is calculated by adding all the mutual information terms between f_1 and each label [17,31]. In summary, the derivation confirms that the calculation for $Rel(f_1)$ in conventional methods can be supported theoretically.

3.3. Objective of designated criterion

After selecting f_1 at the first iteration, the relevance of the remaining features in $\{F - S_1\}$ can be changed because the uncertainty of each label is reduced by f_1 [19]. Thus, the relevance evaluation conditioned by the previously selected feature is required in order to obtain an effective feature subset. With regard to f_1, the dependency of f_2 on L can be measured as

$$\widetilde{M}_1(f_2; L|f_1) = \sum_{l \in L} M(f_2; l|f_1) \tag{10}$$

where $0 \leq M(f_2; l|f_1)$ is the conditional mutual information that measures the dependency of f_2 on l conditioned by f_1, and defined as

$$M(f_2; l|f_1) = H(f_1, f_2) + H(f_1, l) - H(f_1, f_2, l) - H(f_1) \tag{11}$$

Although measuring the conditional merit of candidate features is suitable, employing Eq. (10) in its original form requires repetitive dependency calculations against each label to determine the best feature. Thus, erroneous relevance evaluation can be incurred because of $|L|$ and $|S_{i-1}|$. To remedy this, the algorithm might approximate $M(f_2; l|f_1)$ based on the relationship between two variables [21,22]. Assume that f_2 is mutually dependent on l and mutually independent of f_1, which is the most optimistic case for f_2 being the best candidate. Because $H(f_1, f_2, l) = H(f_1, l)$ and $H(f_1, f_2) = H(f_1) + H(f_2)$ in this case, $M(f_2; l|f_1)$ is simplified to

$$\begin{aligned} M(f_2; l|f_1) &= H(f_1, f_2) + H(f_1, l) - H(f_1, f_2, l) - H(f_1) = H(f_1, f_2) \\ &\quad + H(f_1, l) - H(f_1, l) - H(f_1) = H(f_1, f_2) - H(f_1) = H(f_2) \end{aligned} \tag{12}$$

Eq. (12) indicates that f_2 should be dependent on l and independent of f_1 in parallel to achieve the maximum value $H(f_2)$. In conventional methods, the relevance of f_2 is approximated by subtracting the dependency between f_1 and f_2 from the dependency of f_1 on the labels [21,22]. For example, the relevance evaluation for f_2 at the second step in [21] is

$$\max_{f_2 \in \{F - S_1\}} \left[\sum_{l \in L} M(f_2; l) - M(f_2; f_1) \right] \tag{13}$$

Thus, if L is composed of many labels, the conditional relevance of f_2 can be overestimated. Eq. (13) shows that $Red(f_2)$ is determined without regard to $|L|$, which can vary according to each multi-label dataset. As another example, in the work of [19], the relevance evaluation for f_2 is performed as

$$\max_{f_2 \in \{F - S_1\}} \left[\sum_{l \in L} M(f_2; l) - \sum_{l \in L} I(f_2, f_1, l) \right] \tag{14}$$

As a result, the relevance evaluation incurs the erroneous joint entropy calculation along with l because $I(f_2, f_1, l) = H(f_2) + H(f_1) + H(l) - H(f_2, f_1) - H(f_2, l) - H(f_1, l)$. In

$$+ H(f_2, f_1, l) = M(f_2; f_1) - M(f_2; f_1|l)$$

order to be a scalable relevance evaluation, the inaccurate evaluation should be avoided while circumventing the repetitive calculation. Based on this observation, we propose our relevance evaluation method.

3.4. Scalable criterion for large label set

According to Eq. (1), the relevance evaluation at the second iteration is performed as follows:

$$J = Rel(f_2) - Red(f_2) \tag{15}$$

As Eq. (15) shows, $Rel(f_2)$ and $Red(f_2)$ should be specified to evaluate the relevance of f_2. For the first step, let us start from $Rel(f_2)$. Based on Eq. (9), the dependency of f_2 can be calculated as

$$Rel(f_2) = \sum_{l \in L} M(f_2; l) \tag{16}$$

Next, to grant adaptability to $Red(f_2)$ against the scale of $Rel(f_2)$ and avoid repetitive calculations caused by $f \in S$ and $l \in L$, let $Red(f_2)$ be represented as follows:

$$Red(f_2) = C \cdot Rel(f_2) = C \sum_{l \in L} M(f_2; l) \tag{17}$$

where $0 \le C \le 1$, to be estimated, determines the reduction with relevance to f_2 based on $Rel(f_2)$ while circumventing the repetitive calculations for reduction against each label. As shown in Eq. (12), C should increase if f_2 is dependent on $f_1 \in S_1$. In contrast, it should be close to zero if f_2 is independent of f_1. Thus, this relationship can be emulated using $M(f_2; f_1)$ with an unknown normalization factor to hold the range of C, because $M(f_2; f_1) = 0$ if f_1 and f_2 are independent, but $M(f_2; f_1) = \min(H(f_2), H(f_1))$ if they are dependent. Let C be approximated as follows:

$$C \approx \frac{M(f_2; f_1)}{\alpha} \tag{18}$$

where α is a normalization factor for $M(f_2; f_1)$. Because C should be less than one, the following relationship can be obtained:

$$\frac{M(f_2; f_1)}{\alpha} \le 1 \tag{19}$$

To normalize $M(f_2; f_1)$, α can be chosen from $H(f_2)$ or $H(f_1)$, where each way naturally leads to the coefficient of uncertainty for f_2 and f_1 [4], but its efficacy still has not been investigated for the multi-label feature selection problem. Because $M(f_2; l)$ is upper-bound by $H(f_2)$ and the algorithm focuses on the relevance of f_2 at this iteration, in this paper, $H(f_2)$ is used as the normalization factor. Thus, the relevance of f_2 is adjusted by the portion that represents the dependency on the previously selected feature f_1 against the total information carried by f_2. As a result, C can be rewritten as

$$C \approx \frac{M(f_2; f_1)}{H(f_2)} \tag{20}$$

By combining Eqs. (17) and (20), $R(f_2)$ is rewritten as

$$Red(f_2) \approx \frac{M(f_2; f_1)}{H(f_2)} \sum_{l \in L} M(f_2; l) \tag{21}$$

Thus, the relevance evaluation for f_2 is performed by

$$J = \sum_{l \in L} M(f_2; l) - \frac{M(f_2; f_1)}{H(f_2)} \sum_{l \in L} M(f_2; l) \tag{22}$$

Eq. (22) shows how the relevance of f_2 can be evaluated when the feature subset contains only one feature. Because the number of selected features can be changed, Eq. (22) can be rewritten as

$$J = \sum_{l \in L} M(f_2; l) - \sum_{l \in L} \frac{M(f_2; l)}{H(f_2)} M(f_2; f_1) = \sum_{l \in L} M(f_2; l) - \sum_{f \in S_1} \sum_{l \in L} \frac{M(f_2; l)}{H(f_2)} M(f_2; f) \tag{23}$$

Eq. (23) shows how $Red(f_i)$ is calculated when $i = 2$. By considering the previously selected features in S_{i-1}, the algorithm selects the next feature f_i that maximizes the following condition:

$$\max_{f_i \in \{F - S_{i-1}\}} \left[\sum_{l \in L} M(f_i; l) - \sum_{f \in S_{i-1}} \sum_{l \in L} \frac{M(f_i; l)}{H(f_i)} M(f_i; f) \right] \tag{24}$$

If $H(f_i) = 0$, the algorithm immediately discards f_i during the relevance evaluation process because f_i carries no information. Given a set of previously selected features, the proposed method chooses the next feature f_i that satisfies Eq. (24) in each iteration under incremental selection. For the implementation, it can be beneficial to consider that once $M(f_i; f)$ is calculated, the algorithm does not need to recalculate $M(f_i; l)$ for the calculation of $Red(f_i)$ by iterating all the labels because an equivalent form of Eq. (24) is

$$\max_{f_i \in \{F - S_{i-1}\}} \left[\sum_{l \in L} M(f_i; l) - \sum_{f \in S_{i-1}} \frac{M(f_i; f)}{H(f_i)} \sum_{l \in L} M(f_i; l) \right] \tag{25}$$

where $\sum_{l \in L} M(f_i; l)$ is a constant term for calculating $Red(f_i)$, and because it was already calculated for $Rel(f_i)$, it is reusable.

3.5. Time complexity analysis

Although the main goal of this study is to propose an effective multi-label feature selection method that can improve multi-label learning accuracy within a fixed number of selected features, we conducted an analysis on the proposed method from the perspective of computational cost because this is also an important issue around multi-label feature selection problems. Let k be the number of variables involved in an entropy term. The pattern number is a constant value, and hence the computational cost for calculating an entropy term increases linearly according to k; the number of observations to be examined for calculating entropy is $|W| \cdot k$, where $|W|$ is the number of patterns in a given dataset [17]. For simplicity, we assume a computational cost of $H(X)$, where $|X| = k$ is k unit cost, with one unit cost being the computational cost for calculating an entropy term that involves one variable. Thus, the computational cost for calculating a mutual information term consumes 4 unit costs because $M(a; b) = H(a) - H(a, b) + H(b)$.

Suppose that the proposed method is allowed to use additional memory space to accelerate the relevance evaluation. For clarity, the criterion J for identifying ith feature f_i is written as

$$J = \underbrace{\sum_{l \in L} M(f_i; l)}_{\text{Part 1}} - \sum_{f \in S_{i-1}} \left(\underbrace{M(f_i; f)}_{\text{Part 2}} / \underbrace{H(f_i)}_{\text{Part 3}} \right) \underbrace{\sum_{l \in L} M(f_i; l)}_{\text{Part 4}} \tag{26}$$

Algorithm 1. Proposed method

1: **Input:** n; ▷ Number of features to be selected
2: **Output:** S; ▷ Selected feature subset
3: Initialize $S \leftarrow \{\phi\}$ and $F \leftarrow \{f_1, \ldots, f_d\}$;
4: **repeat**
5: Find the feature $f^+ \in F$ maximizing (25);
6: Set $S \leftarrow \{S \cup f^+\}$, and $F \leftarrow F \backslash S$;
7: **until** $|S| = n$
8: Output the set S containing the selected features;

First, let us focus on the computational cost for Part 1 in Eq. (26). Because Part 1 involves $|L|$ mutual information terms and the calculation for mutual information consumes 4 unit costs, $4 \cdot |L|$ unit costs are consumed to evaluate the relevance of a single feature. Because all the features in F should be evaluated, $4 \cdot |F| \cdot |L|$ unit costs are consumed from the entire feature selection process. Because additional memory usage is allowed, the calculation results of Part 1 can be reused for the calculation of Part 4; the calculation of Part 4 does not incur additional computational costs, and it is thus non-influential on the computational cost. In addition, the calculation of Part 3 is also unnecessary because it is already computed from the calculation of Part 1: $M(f_i; l) = H(f_i) - H(f_i, l) + H(l)$. Next, in each iteration, the algorithm computes a maximum of n mutual information terms for each candidate feature in $\{F - S_i\}$ in order to calculate Part 2. The calculation of each mutual information term incurs 2 unit costs because $M(f_i; f) = H(f_i) - H(f_i, f) + H(f)$, and both $H(f_i)$ and $H(f)$ terms are already calculated from the calculation of Part 1. Thus, the entire feature selection process incurs a maximum of $2 \cdot |F| \cdot n$ unit costs to

evaluate all the candidate features. As a result, the computational cost of the proposed method can be written as $O(4{\cdot}|F|{\cdot}|L| + 2{\cdot}|F|{\cdot}n) = O(|F|{\cdot}|L| + |F|{\cdot}n)$, which is the same computational cost as the works of [21,22], and smaller than the works of [16,19,23]. The detailed steps of the proposed method are described in Algorithm 1.

4. Experimental results

In this section, we demonstrate the performance of the proposed method and other conventional methods in terms of multi-label classification accuracy. To achieve this, we explain about the experimental settings regarding the employed multi-label datasets, multi-label classifier, evaluation measure, and employed statistical tests.

4.1. Experimental settings

We experimented with 25 datasets from different domains. The datasets Bibtex, Delicious, Enron, Language Log (LLog), Slashdot, and RCV1 generated from text mining applications, in which each feature corresponds to the occurrence of a word and each label represents the relevancy of each text pattern to a specific subject. The image dataset Scene is concerned with the semantic indexing of still scenes, in which each scene possibly contains multiple objects. The Medical dataset was sampled from a large corpus of suicide letters obtained after natural language processing of a clinical free text. The Yeast and Genbase datasets resulted from the biological domain, including information about the function of genes and proteins. The remaining 11 datasets were came from the Yahoo dataset collection. We performed an unsupervised dimensionality reduction on the text datasets, such as RCV1 and Yahoo collections, composed of more than 10,000 features. Specifically, the top 2% and 5% features with highest document frequency are retained for RCV1 and Yahoo datasets, respectively [43]. In the text mining domain, the existing studies reported that classification performance will not incur significant changes with the retention of 1% features based on document frequency [40].

Table 1 shows the standard statistics of the multi-label datasets employed in our experiments; these include the number of patterns in the dataset $|W|$, number of features $|F|$, type of features, and number of labels $|L|$. When the feature type is numeric, we discretize the features by using the supervised discretization method [2] for mutual information-based multi-label feature selection methods. Specifically, each observed numeric value is assigned to one of the bins that are automatically determined using the discretization method. The label cardinality *Card* represents the average number of labels for each instance. The label density *Den* represents the label cardinality divided by the total number of labels. The number of distinct label sets *Distinct* indicates the number of unique label subsets in L. *Domain* represents the applications that each dataset is extracted from.

We compared the proposed method with the conventional multi-label feature selection methods: AMI [21], MDMR [23], MLCFS [11], and PPT+RF [30]. For each method, the parameter was set to the value recommended by each research. To choose the number of features to be selected, we referred to the information theory. To distinguish each pattern from the others, we only require $\lceil \log_2|W| \rceil$ features independent to each other even though the type of all features is binary; for example, we need only four binary features to distinguish 16 patterns perfectly. However, this condition can be too strict from real-world situations because features can be dependent to each other. As a result, we choose $\lceil \sqrt{|W|} \rceil$ as the number of features to be selected because it is asymptotically larger than $\log_2|W|$ and implies a fairly small value. To evaluate the quality of a feature subset obtained through each method, we considered the conventional multi-label classifiers [45]. Among them, we chose the multi-label naive Bayes (MLNB) classifier [42] because it outputs the predicted label subset based on the intrinsic characteristics of a given dataset without incurring any complicated parameter-tuning process that may influence the final multi-label classification performance. For fairness, we conducted a holdout cross-validation for each experiment [18]; 80% of the patterns in a given dataset were randomly chosen as the training set for multi-label feature selection and classifier training, and the remaining 20% of the patterns were used as the test set to obtain the multi-label classification performance to be reported. Each experiment was repeated 10 times, and the average value was used to represent the classification performance according to each feature selection method.

We employed three evaluation measures: Hamming loss, Ranking loss, and Normalized coverage. Let $T = \{(T_i, \lambda_i) | 1 \le i \le |T|\}$ be a given test set where $\lambda_i \subseteq L$ is a correct label subset. According to a test

Table 1
Standard characteristics of employed datasets.

Dataset	$\|W\|$	$\|F\|$	Type	$\|L\|$	Card.	Den.	Distinct.	Domain
Bibtex	7,395	1,836	Nominal	159	2.402	0.015	2,856	Text
Delicious	16,105	500	Nominal	983	19.020	0.019	15,806	Text
Enron	1,702	1,001	Nominal	53	3.378	0.064	753	Text
Genbase	662	1,185	Nominal	27	1.252	0.046	32	Biology
LLog	1,460	1,004	Nominal	75	1.180	0.016	304	Text
Medical	978	1,494	Nominal	45	1.245	0.028	94	Text
Scene	2,407	294	Numeric	6	1.074	0.179	15	images
Slashdot	3,782	1,079	Nominal	22	1.181	0.054	156	Text
Yeast	2,417	103	Numeric	14	4.237	0.303	198	Biology
RCV1 (S1)	6,000	945	Numeric	101	2.880	0.029	1,028	Text
RCV1 (S2)	6,000	945	Numeric	101	2.634	0.026	954	Text
RCV1 (S3)	6,000	945	Numeric	101	2.614	0.026	939	Text
RCV1 (S4)	6,000	945	Numeric	101	2.484	0.025	816	Text
RCV1 (S5)	6,000	945	Numeric	101	2.642	0.026	946	Text
Arts	7,484	1,157	Numeric	26	1.654	0.064	599	Text
Business	11,214	1,096	Numeric	30	1.599	0.053	233	Text
Computers	12,444	1,705	Numeric	33	1.507	0.046	428	Text
Education	12,030	1,377	Numeric	33	1.463	0.044	511	Text
Entertain	12,730	1,600	Numeric	21	1.414	0.067	337	Text
Health	9,205	1,530	Numeric	32	1.644	0.051	335	Text
Recreation	12,828	1,516	Numeric	22	1.429	0.065	530	Text
Reference	8,027	1,984	Numeric	33	1.174	0.036	275	Text
Science	6,428	1,859	Numeric	40	1.450	0.036	457	Text
Social	12,111	2,618	Numeric	39	1.279	0.033	361	Text
Society	14,512	1,590	Numeric	27	1.670	0.062	1,054	Text

Table 2
Comparison results of multi-label feature selection methods in terms of Hamming loss (mean ± std. Deviation). The ✓ symbol is attached to the performance value if the corresponding method gives the best performance for each dataset.

Dataset	Proposed	AMI	MDMR	MLCFS	PPT+RF
Bibtex	**0.022 ± 0.001✓**	0.086 ± 0.001	0.085 ± 0.001	0.041 ± 0.003	0.067 ± 0.005
Delicious	0.023 ± 0.001	0.024 ± 0.000	0.024 ± 0.001	**0.020 ± 0.001✓**	0.024 ± 0.006
Enron	**0.063 ± 0.003✓**	0.140 ± 0.003	0.126 ± 0.005	0.069 ± 0.002	0.092 ± 0.009
Genbase	**0.004 ± 0.002✓**	0.010 ± 0.003	0.010 ± 0.003	0.042 ± 0.002	0.010 ± 0.002
LLog	0.022 ± 0.004	0.253 ± 0.010	0.075 ± 0.010	**0.016 ± 0.000✓**	0.095 ± 0.010
Medical	**0.002 ± 0.001✓**	0.018 ± 0.002	0.017 ± 0.003	0.021 ± 0.001	0.011 ± 0.002
Scene	**0.165 ± 0.007✓**	0.256 ± 0.009	0.233 ± 0.005	0.176 ± 0.010	0.198 ± 0.010
Slashdot	**0.042 ± 0.001✓**	0.043 ± 0.002	0.043 ± 0.002	0.044 ± 0.002	0.050 ± 0.002
Yeast	**0.216 ± 0.007✓**	0.276 ± 0.007	0.275 ± 0.007	0.219 ± 0.007	0.236 ± 0.008
RCV1 (S1)	**0.033 ± 0.000✓**	0.056 ± 0.001	0.056 ± 0.001	0.035 ± 0.001	0.040 ± 0.001
RCV1 (S2)	**0.031 ± 0.001✓**	0.060 ± 0.002	0.060 ± 0.002	0.035 ± 0.001	0.038 ± 0.002
RCV1 (S3)	**0.030 ± 0.001✓**	0.059 ± 0.002	0.059 ± 0.002	0.034 ± 0.001	0.039 ± 0.002
RCV1 (S4)	**0.026 ± 0.000✓**	0.057 ± 0.003	0.056 ± 0.003	0.031 ± 0.001	0.035 ± 0.001
RCV1 (S5)	**0.030 ± 0.001✓**	0.062 ± 0.002	0.062 ± 0.002	0.034 ± 0.001	0.044 ± 0.002
Arts	**0.060 ± 0.001✓**	0.072 ± 0.002	0.071 ± 0.002	0.062 ± 0.001	0.086 ± 0.003
Business	0.035 ± 0.001	0.070 ± 0.003	0.068 ± 0.003	**0.035 ± 0.005✓**	0.083 ± 0.003
Computers	**0.043 ± 0.001✓**	0.070 ± 0.003	0.069 ± 0.003	0.045 ± 0.001	0.079 ± 0.006
Education	**0.042 ± 0.001✓**	0.059 ± 0.002	0.058 ± 0.002	0.044 ± 0.001	0.067 ± 0.003
Entertain	**0.059 ± 0.001✓**	0.081 ± 0.002	0.079 ± 0.002	0.062 ± 0.004	0.075 ± 0.002
Health	**0.038 ± 0.001✓**	0.053 ± 0.002	0.053 ± 0.002	0.047 ± 0.004	0.066 ± 0.004
Recreation	**0.056 ± 0.001✓**	0.073 ± 0.003	0.072 ± 0.002	0.057 ± 0.003	0.089 ± 0.003
Reference	**0.031 ± 0.001✓**	0.071 ± 0.004	0.070 ± 0.004	0.034 ± 0.001	0.092 ± 0.004
Science	**0.035 ± 0.002✓**	0.057 ± 0.003	0.055 ± 0.003	0.036 ± 0.002	0.044 ± 0.004
Social	**0.026 ± 0.001✓**	0.052 ± 0.002	0.051 ± 0.002	0.033 ± 0.002	0.087 ± 0.003
Society	**0.053 ± 0.001✓**	0.134 ± 0.007	0.108 ± 0.004	0.056 ± 0.003	0.089 ± 0.003
Avg. Rank	**1.12**	4.52	3.40	2.20	3.76

Table 3
Comparison results of multi-label feature selection methods in terms of Ranking loss (mean ± std. Deviation). The ✓ symbol is attached to the performance value if the corresponding method gives the best performance for each dataset.

Dataset	Proposed	AMI	MDMR	MLCFS	PPT+RF
Bibtex	**0.099 ± 0.004✓**	0.107 ± 0.004	0.105 ± 0.004	0.099 ± 0.006	0.162 ± 0.017
Delicious	**0.116 ± 0.001✓**	0.155 ± 0.004	0.141 ± 0.003	0.119 ± 0.001	0.149 ± 0.008
Enron	**0.096 ± 0.012✓**	0.172 ± 0.014	0.161 ± 0.014	0.100 ± 0.011	0.137 ± 0.012
Genbase	**0.037 ± 0.029✓**	0.038 ± 0.030	0.038 ± 0.030	0.164 ± 0.030	0.040 ± 0.030
LLog	**0.161 ± 0.018✓**	0.178 ± 0.016	0.163 ± 0.019	0.174 ± 0.016	0.217 ± 0.017
Medical	**0.038 ± 0.028✓**	0.044 ± 0.029	0.044 ± 0.029	0.147 ± 0.028	0.047 ± 0.028
Scene	**0.090 ± 0.009✓**	0.164 ± 0.013	0.123 ± 0.010	0.094 ± 0.008	0.141 ± 0.005
Slashdot	**0.210 ± 0.005✓**	0.214 ± 0.006	0.213 ± 0.006	0.221 ± 0.007	0.232 ± 0.007
Yeast	**0.194 ± 0.011✓**	0.249 ± 0.008	0.248 ± 0.008	0.200 ± 0.008	0.208 ± 0.009
RCV1 (S1)	**0.064 ± 0.005✓**	0.073 ± 0.004	0.072 ± 0.004	0.066 ± 0.006	0.083 ± 0.004
RCV1 (S2)	**0.065 ± 0.003✓**	0.070 ± 0.003	0.069 ± 0.003	0.068 ± 0.005	0.086 ± 0.004
RCV1 (S3)	**0.067 ± 0.005✓**	0.077 ± 0.006	0.075 ± 0.005	0.068 ± 0.005	0.092 ± 0.005
RCV1 (S4)	**0.065 ± 0.004✓**	0.077 ± 0.005	0.075 ± 0.005	0.068 ± 0.004	0.081 ± 0.004
RCV1 (S5)	0.066 ± 0.006	0.072 ± 0.006	0.071 ± 0.005	**0.065 ± 0.005✓**	0.079 ± 0.004
Arts	**0.136 ± 0.019✓**	0.148 ± 0.018	0.147 ± 0.018	0.173 ± 0.028	0.185 ± 0.021
Business	**0.059 ± 0.025✓**	0.089 ± 0.024	0.086 ± 0.024	0.061 ± 0.024	0.098 ± 0.025
Computers	**0.079 ± 0.002✓**	0.103 ± 0.003	0.101 ± 0.003	0.090 ± 0.008	0.115 ± 0.003
Education	**0.082 ± 0.002✓**	0.096 ± 0.003	0.095 ± 0.003	0.086 ± 0.003	0.112 ± 0.003
Entertain	**0.105 ± 0.002✓**	0.136 ± 0.003	0.132 ± 0.002	0.128 ± 0.028	0.127 ± 0.004
Health	**0.077 ± 0.028✓**	0.093 ± 0.027	0.091 ± 0.027	0.102 ± 0.033	0.107 ± 0.027
Recreation	**0.129 ± 0.004✓**	0.160 ± 0.004	0.157 ± 0.004	0.148 ± 0.036	0.184 ± 0.005
Reference	**0.096 ± 0.023✓**	0.127 ± 0.022	0.123 ± 0.022	0.103 ± 0.021	0.151 ± 0.023
Science	**0.114 ± 0.005✓**	0.140 ± 0.008	0.133 ± 0.007	0.122 ± 0.008	0.143 ± 0.006
Social	**0.066 ± 0.009✓**	0.087 ± 0.010	0.086 ± 0.010	0.072 ± 0.009	0.104 ± 0.010
Society	**0.126 ± 0.004✓**	0.324 ± 0.007	0.289 ± 0.009	0.140 ± 0.012	0.170 ± 0.005
Avg. Rank	**1.04**	4.00	2.96	2.52	4.48

pattern T_i, a classifier such as MLNB should output a set of confidence values $0 \le \psi_{i,l} \le 1$ for each label $l \in L$. If confidence value $\psi_{i,l}$ is larger than the predefined threshold value, such as 0.5, the corresponding label l can be included in the predicted label subset Y_i. Based on ground truth λ_i, confidence values $\psi_{i,l}$, and predicted label subset Y_i, the multi-label classification performance can be measured according to each evaluation measure [20,43,33]. Hamming loss is defined as

$$hloss(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{|L|}|\lambda_i \triangle Y_i|$$

where $\triangle$ denotes the symmetric difference between two sets. Ranking loss is defined as

$$rloss(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{|\{(a,b)|a \in \lambda_i, b \in \overline{\lambda_i}, \psi_{i,a} \le \psi_{i,b}\}|}{|\lambda_i||\overline{\lambda_i}|}$$

Table 4
Comparison results of multi-label feature selection methods in terms of Normalized coverage (mean ± standard deviation). The ✓ symbol is attached to the performance value if the corresponding method gives the best performance for each dataset.

Dataset	Proposed	AMI	MDMR	MLCFS	PPT+RF
Bibtex	**0.168 ± 0.006✓**	0.180 ± 0.006	0.178 ± 0.007	0.168 ± 0.008	0.249 ± 0.022
Delicious	**0.499 ± 0.005✓**	0.614 ± 0.011	0.574 ± 0.009	0.523 ± 0.004	0.602 ± 0.016
Enron	**0.271 ± 0.012✓**	0.390 ± 0.015	0.378 ± 0.018	0.277 ± 0.010	0.333 ± 0.015
Genbase	**0.084 ± 0.028✓**	0.087 ± 0.029	0.087 ± 0.029	0.214 ± 0.030	0.091 ± 0.029
LLog	**0.205 ± 0.020✓**	0.220 ± 0.018	0.207 ± 0.019	0.216 ± 0.018	0.263 ± 0.017
Medical	**0.069 ± 0.028✓**	0.075 ± 0.030	0.075 ± 0.030	0.180 ± 0.029	0.079 ± 0.028
Scene	**0.257 ± 0.008✓**	0.308 ± 0.011	0.283 ± 0.009	0.261 ± 0.007	0.301 ± 0.005
Slashdot	**0.265 ± 0.006✓**	0.269 ± 0.007	0.268 ± 0.007	0.275 ± 0.009	0.287 ± 0.007
Yeast	0.567 ± 0.014	0.615 ± 0.010	0.615 ± 0.010	**0.558 ± 0.014✓**	0.578 ± 0.009
RCV1 (S1)	**0.150 ± 0.007✓**	0.163 ± 0.005	0.161 ± 0.006	0.152 ± 0.007	0.177 ± 0.005
RCV1 (S2)	**0.150 ± 0.006✓**	0.157 ± 0.006	0.156 ± 0.006	0.151 ± 0.006	0.179 ± 0.006
RCV1 (S3)	0.154 ± 0.005	0.165 ± 0.009	0.162 ± 0.008	**0.153 ± 0.006✓**	0.188 ± 0.006
RCV1 (S4)	0.151 ± 0.006	0.161 ± 0.006	0.158 ± 0.005	**0.149 ± 0.004✓**	0.171 ± 0.005
RCV1 (S5)	0.151 ± 0.007	0.159 ± 0.007	0.157 ± 0.006	**0.148 ± 0.005✓**	0.168 ± 0.004
Arts	**0.236 ± 0.017✓**	0.250 ± 0.017	0.248 ± 0.017	0.270 ± 0.026	0.283 ± 0.019
Business	**0.127 ± 0.024✓**	0.161 ± 0.023	0.158 ± 0.024	0.129 ± 0.024	0.175 ± 0.024
Computers	**0.148 ± 0.002✓**	0.175 ± 0.004	0.172 ± 0.003	0.159 ± 0.009	0.188 ± 0.004
Education	**0.141 ± 0.003✓**	0.158 ± 0.003	0.157 ± 0.003	0.144 ± 0.003	0.172 ± 0.004
Entertain	**0.191 ± 0.003✓**	0.226 ± 0.004	0.222 ± 0.003	0.212 ± 0.025	0.211 ± 0.005
Health	**0.147 ± 0.025✓**	0.163 ± 0.024	0.162 ± 0.024	0.172 ± 0.031	0.180 ± 0.025
Recreation	**0.218 ± 0.005✓**	0.253 ± 0.005	0.250 ± 0.005	0.234 ± 0.036	0.269 ± 0.006
Reference	**0.142 ± 0.022✓**	0.176 ± 0.021	0.172 ± 0.021	0.146 ± 0.021	0.198 ± 0.022
Science	**0.176 ± 0.005✓**	0.205 ± 0.008	0.198 ± 0.007	0.181 ± 0.007	0.206 ± 0.006
Social	**0.116 ± 0.011✓**	0.144 ± 0.011	0.143 ± 0.012	0.121 ± 0.010	0.153 ± 0.011
Society	**0.230 ± 0.004✓**	0.448 ± 0.009	0.411 ± 0.011	0.245 ± 0.013	0.276 ± 0.006
Avg. Rank	**1.16**	4.00	2.96	2.40	4.48

Table 5
Summary of the Friedman statistics F_F (k=5, N=25) and critical value in terms of each evaluation measure.

Evaluation measure	F_F	Critical value (α=0.05)
Hamming loss	62.406	2.466
Ranking loss	63.719	
Normalized coverage	54.370	

where $\overline{\lambda_i}$ is a complementary set of λ_i. Thus, Ranking loss measures the average fraction of (a, b) pairs with $\psi_{i,a} \le \psi_{i,b}$ against all possible relevant and irrelevant label pairs. Finally, Normalized coverage is defined as

$$ncov(T) = \frac{1}{|L|}\left(\frac{1}{|T|}\sum_{i=1}^{|T|}\max_{l\in\lambda_i} rank(l) - 1\right)$$

where $rank(\cdot)$ returns the rank of the corresponding relevant label $l \in \lambda_i$ according to $\psi_{i,l}$ in non-increasing order. Thus, Normalized coverage measures how many labels should be determined as positive for all the relevant labels to be positive. Lower values of Hamming loss, Ranking loss, and Normalized coverage commonly indicate good classification performance.

To analyze the performances of the comparing multi-label feature selection methods, we employed the Friedman test, which is a widely-used statistical test, for comparing multiple methods over a number of datasets [5]. Given k methods and N datasets, let r_i^j denote the rank of the j-th method on the i-th dataset (mean ranks are shared in case of ties). Let $R_j = \frac{1}{N}\sum_{i=1}^{N} r_i^j$ denote the average rank for the j-th method under the null hypothesis (i.e., all the methods have an equal performance); then, the following Friedman statistic F_F will be distributed according to the F-distribution with $k-1$ numerator degrees of freedom and $(k-1)(N-1)$ denominator degrees of freedom:

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

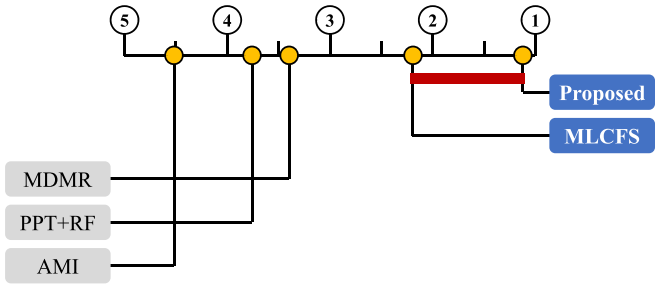
where

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

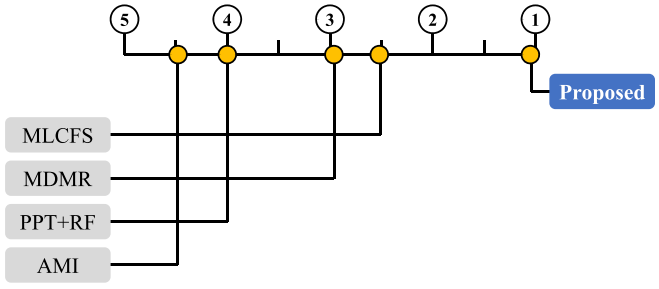
The null hypothesis of equal performance among the comparing algorithms is rejected in terms of each evaluation measure if F_F is larger than the critical value at significance level α. In this case, we need to proceed with a certain post-hoc test to analyze the relative performance among the comparison methods [5]. As we are interested in whether the proposed method achieves better performance than the comparison methods, the Bonferroni–Dunn test is employed [7]. Here, the difference between the average ranks of the proposed method and one comparison method is compared with the following critical difference (CD).

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

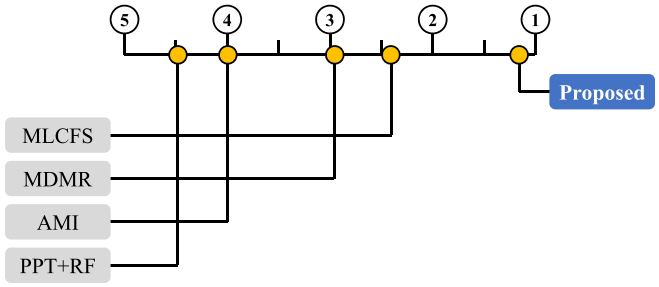
For Bonferroni–Dunn test, the performance between the proposed method and one comparison method is deemed to be statistically similar if their average ranks over all datasets are within one CD. For N=25 and k=5, the critical value at the significance level α=0.05 is 2.466, and the CD with α=0.05 is 1.117 because q_α = 2.498 [5]. In addition, because we are interested in the superiority of the proposed method over conventional mutual information-based multi-label feature selection methods when a large number of labels is involved, we choose 10 multi-label datasets from among 25 datasets based on the number of labels, and then conduct the Wilcoxon signed-rank test [38] to validate the performance of the proposed method in such a situation. Let d_i be the difference between the performance of the two methods on the i-th dataset. The differences are ranked according to their absolute values: the smallest d_i is assigned to the first rank. In the case of ties, average ranks are assigned. Let R^+ be the sum of the ranks for the datasets, on which the compared method outperforms the proposed method, defined as



(a) Hamming loss



(b) Ranking loss



(c) Normalized coverage

Fig. 1. Bonferroni–Dunn test results of five comparison methods with three evaluation measures. Methods not connected with the best method in the CD diagram are considered to show significantly different performance (Significance level α=0.05).

Table 6
Wilcoxon signed-rank test results of the proposed method against other mutual information-based methods for 10 multi-label datasets with large number of labels at the significance level α=0.05 (Sum of outperformed rank R^+ per sum of total rank and p-values shown in the parenthesis).

Evaluation measures	Comparing methods	
	Proposed **versus** AMI	Proposed **versus** MDMR
Hamming loss	**win** (0/55, 1.95e-3)	**win** (0/55, 1.95e-3)
Ranking loss	**win** (0/55, 1.95e-3)	**win** (0/55, 1.95e-3)
Normalized coverage	**win** (0/55, 1.95e-3)	**win** (0/55, 1.95e-3)

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

and R^- be the sum of the ranks for the datasets on which the proposed method outperforms the compared methods. Then, according to the critical values for the Wilcoxon's test, for a confidence level of α=0.05 and N=10, the difference between the compared methods is significant if $\min(R^+, R^-)$ is equal or less than 8. In this case, the null hypothesis of equal performance is rejected.

4.2. Comparison results

Tables 2–4 detail the experimental results of each comparison method on 25 multi-label datasets by using the average performance of holdout cross-validation with the corresponding standard deviation. The best performance among the five comparison methods is represented in bold with a ✓ symbol. In addition, the average rank of each comparison method over all the multi-label datasets is presented in the last row of each table. Table 5 represents the Friedman statistics F_F and the corresponding critical values on each evaluation measure. In addition, at the significance level α=0.05, the null hypothesis of equal performance among the comparing methods is clearly rejected in terms of each evaluation measure.

To show the relative performance of the proposed method and conventional multi-label feature selection methods, Fig. 1 illustrates the CD diagrams on each evaluation measure, where the average rank of each method is marked along the axis placed on the top with better ranks placed on the right-side of each figure. In each figure, any comparison method, with average rank within one CD to that of the best method, is interconnected with a thick line, whose length indicates the extent of one CD on a diagram. Otherwise, any method not connected with the best method is considered to have significantly different performance.

Lastly, Table 6 shows the result of the Wilcoxon signed-rank test of the proposed method against other mutual information-based methods, such as AMI and MDMR, for 10 multi-label datasets with large number of labels: Bibtex, Delicious, Enron, LLog, Medical, and RCV1 datasets: at a significance level α=0.05. For each evaluation measure, the winner of each comparison setting is remarked with bold face, and the corresponding sum of the outperformed rank R^+ per sum of total rank and p-values are presented in the parenthesis.

Based on the empirical experiments and statistical analysis, we can observe several indications. As shown in Tables 2–4, the proposed method outperformed the conventional methods for most multi-label datasets. Specifically, the proposed method achieved the best performance on 88% datasets from Hamming loss experiments, 96% datasets from Ranking loss experiments, and 84% datasets from Normalized coverage experiments. As a result, the proposed method always ranked the best in terms of average rank from these experiments. As determined by the experimental results shown in Fig. 1, the proposed method outperforms the conventional methods. In particular, the proposed method significantly outperforms AMI, MDMR, and PPT+RF, regardless of the evaluation measures. It is interesting to note that the proposed method outperforms other mutual information-based multi-label feature selection methods by a large margin in all the experiments, indicating that the proposed approximation scheme for mutual information between feature subset and given labels is more effective than that of the conventional methods. Furthermore, this is considerably highlighted from the experimental results shown in Table 6 because the proposed method always performed better than AMI and MDMR in terms of three evaluation measures; sum of the outperformed rank R^+ is zero for all the experiments. As these experiments were conducted on 10 selected multi-label datasets with large number of labels, this result indicates that the proposed method does not lose its superiority over conventional mutual information-based methods even though the number of labels is large.

4.3. Analysis on selected features

In our experiments, the proposed method delivers significantly better discriminating power than the other mutual information-based multi-label feature selection methods on several multi-label datasets. To conduct a more detailed analysis, we particularly chose the Bibtex dataset because it is composed of redundant features and fairly large labels among the employed datasets. For example, the 952th and 1824th features of the Bibtex dataset are the same. Thus, one of these

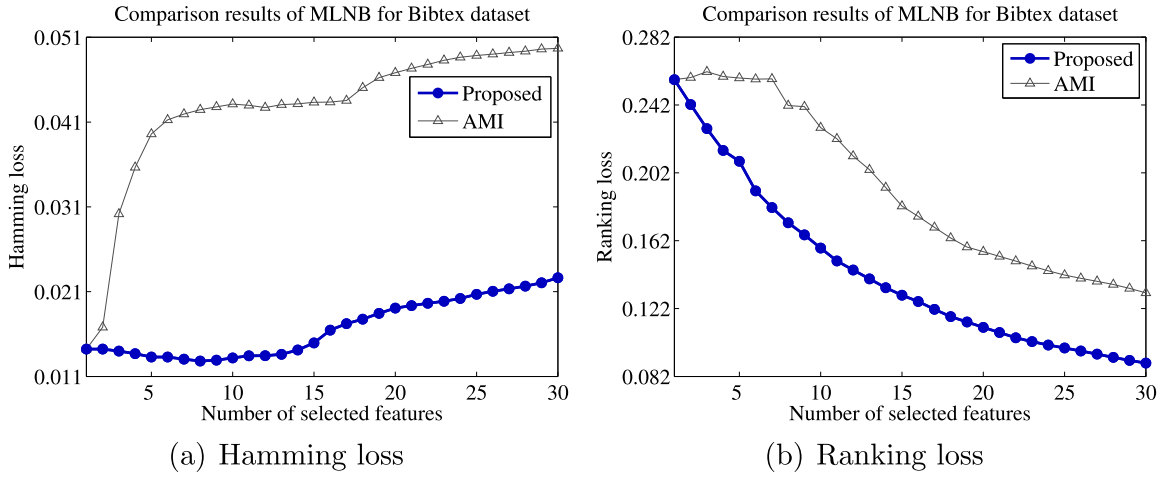


Fig. 2. Hamming and ranking loss performances of Bibtex dataset according to feature subsets using proposed method and AMI.

Table 7
Relevance values calculated using the proposed method at each iteration.

Features	1st iteration	2nd iteration
f_1	0.199	0.198
⋮	⋮	⋮
f_{382}	1.007	**0.633**
⋮	⋮	⋮
f_{952}	**1.596**	–
⋮	⋮	⋮
f_{1824}	1.596	0.000
⋮	⋮	⋮
Feature subset	$\{f_{952}\}$	$\{f_{952}, f_{382}\}$

Table 8
Relevance values calculated using the compared method at each iteration.

Features	1st iteration	2nd iteration
f_1	0.199	0.197
⋮	⋮	⋮
f_{382}	1.007	0.693
⋮	⋮	⋮
f_{952}	**1.596**	–
⋮	⋮	⋮
f_{1824}	1.596	**1.009**
⋮	⋮	⋮
Feature subset	$\{f_{952}\}$	$\{f_{952}, f_{1824}\}$

Table 9
Values required for calculating the relevance of each feature.

Entropy	$H(f_{382})$	$H(f_{952})$	$H(f_{1824})$
Values	0.845	0.587	0.587
Mutual Information	$M(f_{952}; f_{382})$	$M(f_{952}; f_{1824})$	
Values	0.314	0.587	

Table 10
Calculated result of D and R for proposed Method and compared method in the second iteration.

Method	$Rel(f_{1824})$	$Red(f_{1824}, f_{952})$	Relevance
Proposed	1.596	1.596	0.000
AMI	1.596	0.587	1.009

features should not be included into the final feature subset during the selection of the other feature so as to have a compact feature subset. To perform a direct comparison, we chose AMI [21] as the counterpart for the proposed method because *Rel* on both methods is the same, and only *Red* is calculated differently. In particular, we investigated the composition of the feature subset in each method by analyzing certain selected features from each iteration. For brevity, in this section, the subscript of a feature represents the actual index. For example, f_{952} indicates the 952th feature in the Bibtex dataset.

Fig. 2 shows the Hamming and Ranking loss performances of Bibtex dataset according to feature subsets using the proposed method and AMI. In each figure, the horizontal and vertical axes respectively represent the number of selected features and corresponding multi-label classification performances from the perspective of each evaluation measure. The experimental results indicate that the proposed method significantly outperforms AMI, regardless of the number of selected features. More importantly, the results show a common observation: the second feature selected by AMI cannot improve the classification performance, or leads to an inferior classification performance. To identify the cause of this observation, we represent the relevance values calculated by the proposed method in each iteration from Table 7. The first column represents each of the 1,836 features in the Bibtex dataset sequentially. The second column comprises the corresponding relevance value of each feature in the first iteration. As f_{952} provides the best relevance value, this feature is included in the S feature subset in the first iteration. Next, in the second iteration, the relevance value of each feature is reduced because of the previously selected feature f_{952}, resulting in the selection of f_{382} in the current iteration, and $S = \{f_{952}, f_{382}\}$. For comparison, we represent the relevance values calculated using AMI at each iteration in Table 8. As the method for the relevance calculation is the same for both the proposed method and AMI in the first iteration, the same feature subset is obtained in the first iteration. However, in the second iteration, the proposed method and AMI select different features because the relevance evaluation under the selection of f_{382} is different: AMI identifies f_{1824} as the best candidate for the second iteration. It is interesting to note that the relevance value for f_{1824} calculated by the proposed method is zero, indicating a significantly different preference against the features between the two methods. As aforementioned, the features f_{952} and f_{1824} are the same. Thus, during the selection of f_{952}, f_{1824} never carries any additional discriminating power to S. Therefore, the relevance of f_{1824} should be evaluated as a low value, such as zero. In other words, AMI is assigned an excessively large relevance to the unnecessary feature and f_{1824} is then chosen as the best feature for the second iteration.

In our experiments, Ranking loss performance on the Bibtex

dataset improved consistently with the proposed method, as shown in Fig. 2(b), whereas the second feature selected by AMI failed to improve the classification performance. Because a feature that is dependent on the previously selected feature does not significantly change the classification performance, this indicates that the features selected by AMI might be dependent on the previously selected feature. To validate this expectation, we conducted a final analysis on the features selected by the proposed method and AMI in the second iteration. Table 9 demonstrates the required entropy and mutual information values for the relevance evaluation in each method. The table indicates that f_{1824} is mutually dependent on f_{952} because $H(f_{952}) = H(f_{1824}) = M(f_{952}; f_{1824})$, indicating that f_{1824} does not carry any additional discriminating power under the selection of f_{952}. Based on these values, we represent the calculation results of *Rel* and *Red* for f_{1824} in Table 10. As the table indicates, f_{1824} is discarded by the proposed method because the relevance is evaluated as zero, given that it does not carry additional information after f_{952} is included in S.

5. Discussion

In real-world applications, the dataset may include irrelevant or redundant features because of the lack of prior knowledge. Theoretically, a pattern may lose its individuality because of such features because the similarity of each pair of patterns in the same class can be decreased [37], leading to confusion in the learning algorithm and poor classification performance. We can infer that the impact of such features becomes more severe in multi-label learning situations because this adverse effect concurrently occurs for multiple labels. As the importance of features is determined using the score function of a feature selection method, it is important to develop a score function that is able to imply an accurate score of given features. Otherwise, the classification performance can degrade even though an additional feature is included in the feature subset because of inaccurate score values, as shown in Fig. 2(a). A usual approach to evaluate the score of features is by calculating the dependency between candidate feature subsets and all the labels. However, because the number of training patterns is limited, the exact dependency calculation can be an impractical task when the size of candidate feature subset is large. This problem may be avoided by employing the incremental selection strategy. As the incremental selection strategy creates the final feature subset by starting from an empty feature subset and adding a feature at each iteration, the score function will inherently evaluates the dependency of feature subsets composed of small number of features. Thus, a problem on the feature side can be circumvented effectively [19].

In contrast to the single-label feature selection, for the multi-label feature selection, a problem still exists on the label side: an explosive number of distinct label subsets. For example, the Bibtex dataset is composed of 7,395 patterns that are assigned to 2,856 distinct label subsets. This indicates that the full order approach, which considers each distinct label subset as a class, will be unfavorable when a large number of labels are involved. As a result, many multi-label feature selection studies focus on developing a method to measure the dependency based on first- [21,22] or second-order [16,17,19,23] relations between features and labels, leading to a series of different approximation methods. One advantage of second-order approach over first-order approach is the calculation of dependency between feature-pairs from the viewpoint of each label, leading to the sensitive relevance evaluation of features. As a result, the second-order approach, such as MDMR, can outperform the first-order approach, such as AMI, as shown in Fig. 1. However, when a large number of labels are involved, the multi-label feature selection method must consider several relations among feature–feature–label combinations for performing the relevance evaluation that cannot be remedied by the incremental selection strategy. Thus, we can expect that the risk of inaccurate relevance evaluation increases according to the number of labels because the error, that is, the difference between the estimated and true probabilities from the calculation for each combination or relation, will be cumulated to the final relevance value.

The main contribution of this study is the proposal of an effective multi-label feature selection method based on a new approximation for the relevance evaluation based on the first-order approach. Unlike our previous studies [16,17,19], the proposed approximation does not rely on the second-order relations among features and labels to remedy the possible risk of cumulative error from the probability estimation of a large number of labels. Instead, we propose a scalable approximation method based on the first-order approach, considering a large number of labels. To achieve this, we devised a way to approximate the dependency of feature-pairs in terms of each label without incurring the calculation for the second-order relations. As such, the proposed method solves the problem of the first-order based previous studies [21,22] that may choose a feature carrying no additional discriminating power when a large number of labels is involved, as shown in Section 4.3, and statistically superior classification performance than the second-order approaches, as shown in Section 4.2.

6. Conclusion

We proposed a new multi-label feature selection method based on scalable relevance evaluation. To identify the cause of inaccurate relevance evaluation in the conventional multi-label feature selection methods, the characteristics of relevance evaluation were demonstrated theoretically. In order to perform accurate relevance evaluation against dependency on multiple labels, an effective approximation for measuring feature dependency under a multi-label situation was employed and modified for the multi-label feature selection problem. The results showed that the proposed method was able to evaluate the relevance of the features more accurately than the conventional methods, resulting in improvements to multi-label learning accuracy.

Future research direction will be the search method, which is another fundamental building block in the multi-label feature selection method. Although the experimental results showed that the feature subset selected by the proposed method delivers significantly superior multi-label learning accuracy compared with the conventional methods, the multi-label learning accuracy could be improved further by employing a different search method. We would like to study this issue further.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2016R1C1B1014774).

References

[1] R. Agrawal, A. Gupta, Y. Prabhu, M. Varma, Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages, in: Proceedings of the 22nd International Conference World Wide Web, Rio de Janeiro, Brazil, 2013, pp. 13–24.
[2] A. Cano, J.M. Luna, E.L. Gibaja, S. Ventura, LAIM discretization for multi-label data, Inf. Sci. 330 (2016) 370–384.
[3] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, MLSMOTE: approaching imbalanced multilabel learning through synthetic instance generation, Knowl. -Based Syst. 89 (2015) 385–397.
[4] T. Cover, J. Thomas, J. Wiley, et al., Elements of Information Theory 6, Wiley Online Library, 1991.
[5] J. Demsar, Statistical comparisons of classifier over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[6] G. Doquire, M. Verleysen, Mutual information-based feature selection for multi-label classification, Neurocomputing 122 (2013) 148–155.
[7] O.J. Dunn, Multiple comparisons among means, J. Am. Stat. Assoc. 56 (1961) 52–64.
[8] P. Duygulu, K. Barnard, J.F. de Freitas, D.A. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, in: Proceedings of the 7th European Conference Computer Vision, Copenhagen,

Denmark, 2002, pp. 97–112.
[9] Q. Gu, Z. Li, J. Han, Correlated multi-label feature selection, in: Proceedings of the 20th ACM International Conference Information and Knowledge Management, Glasgow, UK, 2011, pp. 1087–1096.
[10] S. Ji, J. Ye, Linear dimensionality reduction for multi-label classification, in: Proceedings 21th International Joint Conference Artificial Intelligence, Pasadena, USA, 2009, pp. 1077–1082.
[11] S. Jungjit, M. Michaelis, A.A. Freitas, J. Cinatl, Two extensions to multi-label correlation-based feature selection: A case study in bioinformatics, in: IEEE International Conference Systems, Man, and Cybernetics, Manchester, UK, 2013, pp. 1519–1524.
[12] I. Katakis, G. Tsoumakas, I. Vlahavas, Multilabel text classification for automated tag suggestion, ECML PKDD Discov. Chall. 75 (2008).
[13] D. Kong, C. Ding, H. Huang, H. Zhao, Multi-label ReliefF and F-statistic feature selections for image annotation, in: Proceedings IEEE Conference Computer Vision and Pattern Recognition, Providence, USA, 2012, pp. 2352–2359.
[14] X. Kong, P. Yu, gMLC: a multi-label feature selection framework for graph classification, Knowl. Inf. Syst. 31 (2) (2012) 281–305.
[15] N. Kwak, C.H. Choi, Input feature selection for classification problems, IEEE Trans. Neural Net. 13 (2002) 143–159.
[16] J. Lee, D.W. Kim, Feature selection for multi-label classification using multivariate mutual information, Pattern Recognit. Lett. 34 (2013) 349–357.
[17] J. Lee, D.W. Kim, Fast multi-label feature selection based on information-theoretic feature ranking, Pattern Recognit. 48 (2015) 2761–2771.
[18] J. Lee, D.W. Kim, Memetic feature selection algorithm for multi-label classification, Inf. Sci. 293 (2015) 80–96.
[19] J. Lee, D.W. Kim, Mutual information-based multi-label feature selection using interaction information, Expert Syst. Appl. 42 (2015) 2013–2025.
[20] J. Lee, H. Kim, N.R. Kim, J.H. Lee, An approach for multi-label classification by directed acyclic graph with label correlation maximization, Inf. Sci. 351 (2016) 101–114.
[21] J. Lee, H. Lim, D.W. Kim, Approximating mutual information for multi-label feature selection, Electron. Lett. 48 (15) (2012) 929–930.
[22] H. Lim, J. Lee, D.W. Kim, Multi-label learning using mathematical programming, IEICE Trans. Inf. Syst. 98 (2015) 197–200.
[23] Y. Lin, Q. Hu, J. Liu, J. Duan, Multi-label feature selection based on max-dependency and min-redundancy, Neurocomputing 168 (2015) 92–103.
[24] Y.C. Lin, Y.H. Yang, H.H. Chen, Exploiting online music tags for music emotion classification, ACM Trans. Multimed. Comput. Commun. Appl. 7 (2011) 26.
[25] S.M. Liu, J.H. Chen, A multi-label classification based approach for sentiment classification, Expert Syst. Appl. 42 (2015) 1083–1093.
[26] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint $l_{2,1}$-norms minimization, Adv. Neural Inf. Process. Syst. 23 (2010) 1813–1821.
[27] B. Qian, I. Davidson, Semi-supervised dimension reduction for multi-label classification, in: Proceedings of the 24th AAAI Conference Artificial Intelligence, Atlanta, USA, 2010, pp. 569–574.
[28] Y. Rao, Contextual sentiment topic model for adaptive social emotion classification, IEEE Intell. Syst. 31 (2016) 41–47.
[29] J. Read, A pruned problem transformation method for multi-label classification, in: Proceedings New Zealand Computer Science Research Student Conference, Christchurch, New Zealand, 2008, pp. 143–150.
[30] 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.
[31] C. Shannon, A mathematical theory of communication, ACM SIGMOBILE Mob. Comput. Commun. Rev. 5 (2001) 3–55.
[32] N. Spolaôr, E.A. Cherman, M.C. Monard, H.D. Lee, A comparison of multi-label feature selection methods using the problem transformation approach, Electron. Notes Theor. Comput. Sci. 292 (2013) 135–151.
[33] N. Spolaôr, M.C. Monard, G. Tsoumakas, H.D. Lee, A systematic review of multi-label feature selection and a new method based on label construction, Neurocomputing 180 (2016) 3–15.
[34] Y. Sun, A. Wong, M. Kamel, Classification of imbalanced data: a review, Int. J. Pattern Recognit. Artif. Intell. 23 (2009) 687–719.
[35] K. Trohidis, G. Tsoumakas, G. Kalliris, I. Vlahavas, Multi-label classification of music into emotions, in: Proceedings 9th International Society Music Information Retrieval, Philadelphia, USA, 2008, pp. 325–330.
[36] G. Tsoumakas, I. Katakis, I. Vlahavas, Random k-labelsets for multilabel classification, IEEE Trans. Knowl. Data Eng. 23 (7) (2011) 1079–1089.
[37] S. Watanabe, Knowing and Guessing a Quantitative Study of Inference and Information, Wiley, New York, 1969.
[38] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (1945) 80–83.
[39] B. Wu, E. Zhong, A. Horner, Q. Yang, Music emotion recognition by multi-label multi-layer multi-instance multi-view learning, in: Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, USA, 2014, pp. 117–126.
[40] Y. Yang, S. Gopal, Multilabel classification with meta-level features in a learning-to-rank framework, Mach. Learn. 88 (2012) 47–68.
[41] Y.H. Yang, H.H. Chen, Machine recognition of music emotion: a review, ACM Trans. Intell. Syst. Technol. 40 (2012) 1–40.
[42] M.L. Zhang, J.M. Peña, V. Robles, Feature selection for multi-label naive bayes classification, Inf. Sci. 179 (2009) 3218–3229.
[43] M.L. Zhang, L. Wu, LIFT: multi-label learning with label-specific features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 107–120.
[44] M.L. Zhang, Z.H. Zhou, ML-kNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (2007) 2038–2048.
[45] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (2014) 1819–1837.

Jaesung Lee takes Post-doctoral course at Chung-Ang University in Seoul, Korea, in the School of Computer Science and Engineering. He is currently interested in data mining with applications to biomedical informatics and affective computing. In theoretical domain, he also studies classification, feature selection, and especially multi-label learning with information theory.

Dae-Won Kim is currently an associated professor in the School of Computer Science and Engineering, Chung-Ang University in Seoul, Korea. Prior to coming to Chung-Ang University, he did his postdoc, Ph.D., M.S. at KAIST, and the B.S. at Kyungpook National University, Korea. His research interest includes advanced data mining algorithms with innovative applications to bioinformatics, music emotion recognition, educational data mining, affective computing, and robot interaction.