Hindawi
Complexity
Volume 2018, Article ID 6292143, 15 pages
https://doi.org/10.1155/2018/6292143

WILEY | Hindawi

Research Article

Scalable Multilabel Learning Based on Feature and Label Dimensionality Reduction

Jaesung Lee[1] **and Dae-Won Kim**[2]

[1]*Chung-Ang University, Seoul, Republic of Korea*
[2]*The School of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea*

Correspondence should be addressed to Dae-Won Kim; dwkim@cau.ac.kr

Received 3 January 2018; Revised 15 July 2018; Accepted 16 August 2018; Published 24 September 2018

Academic Editor: Ireneusz Czarnowski

The data-driven management of real-life systems based on a trained model, which in turn is based on the data gathered from its daily usage, has attracted a lot of attention because it realizes scalable control for large-scale and complex systems. To obtain a model within an acceptable computational cost that is restricted by practical constraints, the learning algorithm may need to identify essential data that carries important knowledge on the relation between the observed features representing the measurement value and labels encoding the multiple target concepts. This results in an increased computational burden owing to the concurrent learning of multiple labels. A straightforward approach to address this issue is feature selection; however, it may be insufficient to satisfy the practical constraints because the computational cost for feature selection can be impractical when the number of labels is large. In this study, we propose an efficient multilabel feature selection method to achieve scalable multilabel learning when the number of labels is large. The empirical experiments on several multilabel datasets show that the multilabel learning process can be boosted without deteriorating the discriminating power of the multilabel classifier.

1. Introduction

Nowadays, the data-driven management of real-life systems based on a model obtained by analyzing data gathered from its daily usage is attracting significant attention because it realizes scalable control for large-scale and complex systems [1, 2]. Unfortunately, advances in the identification of important knowledge on the relation between the observed information and target concept are far from satisfactory for real-life applications such as text categorization [3], protein function prediction [4], emotion recognition [5], and assembly line monitoring [6]. This is because the underlying combinatorial optimization problem is computationally difficult. To deal with this complicated task in a scalable manner, the algorithm may need to identify essential data that carries important knowledge for building an acceptable model while satisfying practical constraints such as real-time response, limited data storage, and computational capability [7].

Although the majority of current machine learning algorithms are designed to learn the relation between information sources or features and a single concept or label, recent complex applications require that the algorithm extracts the relation to multiple concepts [8]. For example, a document can be assigned to multiple categories simultaneously [9], and protein compounds can also have multiple roles in a biological system [10]. Therefore, to identify important knowledge in this scenario, the algorithm must learn the complex relation between features and labels, formalized as the multilabel learning problem in this field. This scenario differs from that of the single-label learning problem because the problem itself offers the opportunity to improve learning accuracy by exploiting the dependency between labels [11, 12]. However, the algorithm eventually suffers as a result of the computational cost of the learning process owing to the multiple labels.

To reduce the computational burden of the algorithm, a straightforward approach is to ignore unimportant features in the training process that do not influence the learning quality [13, 14]. However, in the multilabel learning problem, this approach may be insufficient to satisfy the practical

constraints because a large number of labels can be involved in a related application. Moreover, the possible combinations of features and labels that should be considered for scoring the importance of features increases exponentially; i.e., the feature selection process can become computationally impractical. Additionally, the computational burden increases significantly because the number of features in the dataset is typically large when feature selection is considered. As a result, the number of possible combinations can increase considerably [15]. This is a serious problem because conventional multilabel learning algorithms with and without the feature selection process are unable to finish the learning process owing to the presence of too many features and the scoring process of the features, respectively.

In this study, we devise a new multilabel feature selection method that facilitates dimensionality reduction of labels from the scoring process. Specifically, our algorithm first analyzes the amount of information content in labels and reduces the computational burden by discarding labels that are unimportant to the scoring of the importance of features. Our contribution to this study compared to our previous works and the strategy to deal with the scalability issue can be summarized as follows:

(i) We propose an efficient multilabel feature selection method based on the simplest approximation of mutual information (MI) that is scalable to the number of labels; it costs constant time computations in terms of the number of labels

(ii) The computational cost of the feature selection process can be controlled easily owing to its simple form. This is an important property when the execution time is limited

(iii) The proposed method identifies a subset of labels that carries the majority of the information content compared to the original label set to preserve the quality of the scoring process

(iv) According to the characteristics of labels in terms of information content, we suggest that the size of labels be considered in the feature scoring process to preserve the majority of the information content

(v) In contrast to our previous works, the proposed method explicitly discards unimportant labels from the scoring process, resulting in a significant acceleration of the multilabel feature selection process

2. Multilabel Feature Selection

One of the most common methods of multilabel feature selection is the use of the conventional single-label feature selection method after transforming label sets into one or more labels [9, 16, 17]. In this regard, the simplest strategy is known as binary relevance, in which each label is separated and analyzed independently [18]. A statistical measure that can be used as a score function to measure feature importance can be employed after separating the label set; these measures include the Pearson correlation coefficient [19] and the odds ratio [20]. Thus, prohibitive computations may be required to obtain the final feature score if a large label set is involved. In contrast, efficient multilabel feature selection may not be achieved if the transformation process consumes excessive computational resources. For example, ELA + CHI evaluates the importance of each feature using χ^2 statistics (CHI) between the feature and a single label obtained by using entropy-based label assignment (ELA), which separates multiple labels and assigns them to duplicated patterns [9]. Thus, the label transformation process can be the bottleneck that incurs a prohibitive execution time if the multilabel dataset is composed of a large number of patterns and labels.

Although the computational cost of the transformation process can be reduced by applying a simple procedure such as a label powerset that treats each distinct label set as a class [17, 21], the feature selection process may be inefficient if the scoring process incurs excessive computational costs during the evaluation of the importance of the features [18, 22]. For example, PPT + RF identifies appropriate weight values for the features based on a label that is transformed by the pruned problem transformation (PPT) [21] and the conventional ReliefF (RF) scheme [23] for single-label feature selection [24]. Although the ReliefF method can be extended to handle multilabel problems directly [25], the execution time to obtain the final feature subset can be excessively long if the dataset is composed of a large number of patterns. This is because ReliefF requires similarity calculations for pattern pairs. Thus, the feature selection process itself should not incur a complicated scoring process to achieve efficient multilabel learning.

Instead of a label set transformation approach that may incur side effects [26], an algorithm adaptation approach that attempts to handle the problem of multilabel feature selection directly is considered [15, 27–31]. In this approach, a feature subset is obtained by optimizing a specific criterion such as a joint learning criterion involving feature selection and multilabel learning concurrently [32, 33], $l_{2,1}$-norm function optimization [31], a Hilbert–Schmidt independence criterion [28], label ranking errors [27], F-statistics [34], label-specific feature selection [12], and memetic feature selection based on mutual information (MI) [35]. However, if multilabel feature selection methods based on this strategy consider all features and labels simultaneously, the scoring process can be computationally prohibitive or even fail owing to the internal task of finding an appropriate hyperspace using pairwise pattern comparisons [27], a dependency matrix calculation [28], and iterative matrix inverse operations [31].

In our previous work [29], we demonstrated that MI can be decomposed into a sum of dependencies between variable subsets, which is a very useful property for solving multilabel learning problems [12, 15] because unnecessary computations can be determined prior to the actual computation and be rejected [36]. More efficient score functions, specialized into an incremental search strategy [37] and a quadratic programming framework [38], have also been considered. These score functions were employed to improve the

effectiveness of evolutionary searching [35, 39]. However, these MI-based score functions commonly require the calculation of the dependencies between all variable pairs composed of a feature and a label [14]. Thus, they share the same drawback in terms of computational efficiency because labels known to have no influence on the evaluation of feature importance are included in the calculations [15, 40]. In contrast to our previous study, our method proposed in this study discards unimportant labels explicitly prior to any multilabel learning process.

Although the characteristics of multilabel feature selection methods can vary according to the manner in which the importance of features is modeled, conventional methods create a feature subset by scoring the importance of features either to all labels [9, 17, 28] or to all possible combinations drawn from the label set [15, 27, 29]. Thus, these methods inherently suffer from prohibitive computational costs when the dataset is composed of a large number of labels.

3. Proposed Method

In this section, a formal definition of the multilabel classification and feature selection is provided. Based on our definition, the proposed label selection approach is described and a discussion on the influences of label subset selection to the feature selection is presented.

3.1. Problem Definition. Let $\mathscr{W}$ be a set of training examples or patterns where each example $w_i \in \mathscr{W} (1 \le i \le |\mathscr{W}|)$ is described by a set of features $\mathscr{F} = \{f_1, \dots, f_{|\mathscr{F}|}\}$; its association to multiple concepts can be represented using a subset of labels $\lambda_i \subseteq \mathscr{L}$, where $\mathscr{L} = \{l_1, l_2, \dots, l_{|\mathscr{L}|}\}$. In addition, let $\mathscr{T} = \{(t_i, \lambda_i) \mid 1 \le i \le |\mathscr{T}|\}$ be a set of test patterns, where λ_i is a true label set for t_i and is unknown to the multilabel classifier, resulting in $\mathscr{U} = \mathscr{W} \cup \mathscr{T}$ and $\mathscr{W} \cap \mathscr{T} = 0$. The task of multilabel learning is to derive a family of $|\mathscr{L}|$ functions, namely, $h_1, h_2, \dots, h_{|\mathscr{L}|}$ that are induced from the training examples, where each function $h_k : t_i \to \mathbb{R}$ outputs the class membership of t_i to l_k. Thus, relevant labels of t_i based on each function can be denoted as $\widehat{\lambda}_i = \{l_k \mid h_k(t_i) > \phi, 1 \le k \le |\mathscr{L}|\}$, where ϕ is a predefined threshold. For example, in the work of [41], a mapping function h_k for l_k is induced using $\mathscr{W}$. Based on h_k, the class membership value $h_k(t_i)$ for the given test pattern t_i is determined, where $h_k(t_i) \in [0, 1]$. In this work, the threshold ϕ is set to 0.5 according to the maximum a posteriori theorem. Although the algorithm outputs l_k as a relevant label for t_i if the class membership value is larger than 0.5 in their work, the range of class membership value can be different according to the multilabel classification algorithm. Although there are some trials to improve the multilabel learning performance by adapting threshold for each label [42], most conventional studies have employed the same value for all the labels.

One of the problems of multilabel feature selection that distinguishes it from classical single-label feature selection is the computational cost for selecting a subset of features with regard to the given multiple labels. The multilabel feature selection can then be achieved through a ranking process by assessing the importance of $|\mathscr{F}|$ features based on a score function and selecting the top-ranked n features from $\mathscr{F} (n \ll |\mathscr{F}|)$. To perform multilabel feature selection, an algorithm must be able to measure the dependency, i.e., importance score, between each feature and label set. The dependency between a feature $f \in \mathscr{F}$ and label set $\mathscr{L}$ can be measured using MI [43].

$$M(f ; \mathscr{L}) = H(f) - H(f, \mathscr{L}) + H(\mathscr{L}), \tag{1}$$

where $H(\cdot)$ of (1) represents a joint entropy that measures the information content carried by given a set of variables, defined as

$$H(X) = -\sum_{x \in X} P(x) \log_a P(x), \tag{2}$$

where x is a state represented by a variable X and $P(\cdot)$ is a probability mass function. If the base of the log function, a in (2), is two, this is known as Shannon entropy. When $|\mathscr{L}|$ is large, the calculation of $H(f, \mathscr{L})$ and $H(\mathscr{L})$ becomes unreliable because of too many joint states coming from $\mathscr{L}$ with insufficient patterns. For example, to observe all the possible associations between patterns and label subsets, the dataset should contain at least $2^{|\mathscr{L}|}$ patterns. Let X^* be the power set of X and $X_k^* = \{e \mid e \in X^*, |e| = k\}$. Equation (1) can then be rewritten using the work of Lee and Kim [15].

$$M(f ; \mathscr{L}) = \sum_{k=2}^{|\mathscr{L}|+1} (-1)^k V_k(f \times \mathscr{L}_{k-1}^*), \tag{3}$$

where $\times$ denotes the Cartesian product of two sets. Next, $V_k(\cdot)$ is defined as

$$V_k(Y) = \sum_{X \in Y_k^*} I(X), \tag{4}$$

where $I(X)$ is the interaction information for a given variable set X, defined as [44]

$$I(X) = -\sum_{Y \in X^*} (-1)^{|Y|} H(Y). \tag{5}$$

Equation (3) indicates that $M(f ; \mathscr{L})$ can be approximated into interaction information terms involving a feature and all the possible label subsets. With regard to (3), the most efficient approximation of (1) is known as [36]

$$\begin{aligned} \tilde{M}(f ; \mathscr{L}) &= V_2(f \times \mathscr{L}_1^*) \\ &= \sum_{X \in \{f \times \mathscr{L}_1^*\}_2^*} I(X) \\ &= \sum_{l \in \mathscr{L}} I(f, l) \\ &= \sum_{l \in \mathscr{L}} M(f ; l). \end{aligned} \tag{6}$$

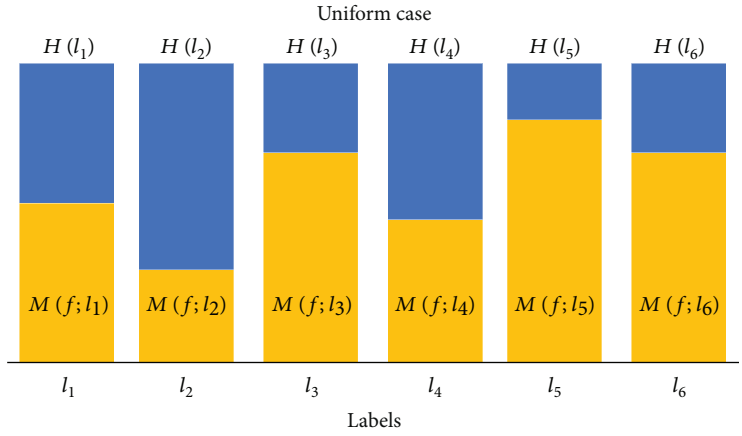


Figure 1: Score value calculation when label entropy values are uniform.

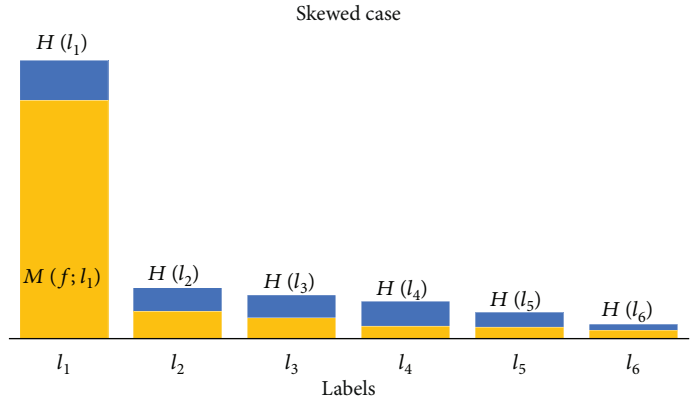


Figure 2: Score value calculation when label entropy values are skewed.

Accordingly, the score function J for evaluating the importance of a given feature f is written as

$$J = \sum_{l \in \mathscr{L}} M(f; l). \tag{7}$$

Equation (7) indicates that the computational cost increases linearly according to $|\mathscr{L}|$. By assuming that the computational cost for calculating a $M(\cdot;\cdot)$ term is a unit cost, the algorithm will consume $|\mathscr{L}|$ unit costs to compute the importance of a feature.

3.2. Label Subset Selection. In our multilabel feature selection problem, the rank of each feature is determined based on importance score using (7). The bound of a MI term is known as

$$0 \le M(f; l) \le \min\left(H(f), H(l)\right). \tag{8}$$

Thus, the bound of (7) is

$$0 \le J \le \sum_{l \in \mathscr{L}} \min(H(f), H(l)). \tag{9}$$

Because $H(f)$ is unknown before actually examining input features and any importance score cannot exceed the sum of entropy value of each label, (9) can be simplified as

$$0 \le J \le \sum_{l \in \mathscr{L}} H(l). \tag{10}$$

Equation (10) indicates that the score value of each feature is influenced by the entropy value of each label, and this fact implies Proposition 1 as follows [40].

Proposition 1 *(upper bound of J). If $\mathscr{L}$ is a given label set, then the upper-bound of J is*

$$\sum_{l \in \mathscr{L}} H(l). \tag{11}$$

Figure 1 represents how the importance score of a feature is determined with regard to Proposition 1; the height of the blue bar indicates the entropy value of the corresponding label, and height of the yellow bar indicates the MI between f and each label. Figures 1 and 2 represent two sample cases wherein each label carries the same amount of information content, and a small subset of label set carries the majority

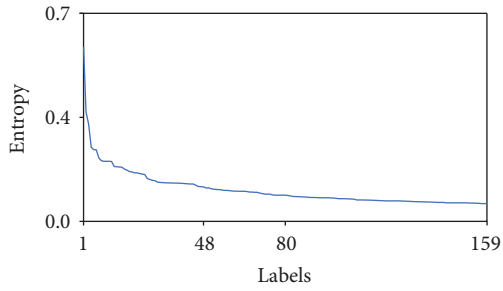


Figure 3: Entropy of each label in BibTeX dataset.

information content, respectively. As shown in Figure 1, the value of $M(f; l_i)$ can be varied according to $l_i \in \mathscr{L}$; however, its value is smaller than the entropy value of each label. When the entropy values of labels are uniformly distributed, all the MI terms between f and each label should be examined because each $M(f; l_i)$ term has same chance of giving significant contribution to the final score J. However, as shown in Figure 2, if there is a set of labels having a small entropy, i.e., if the entropy values of the labels are skewed, there can be MI terms that insignificantly contribute to the extent of J, because all the $M(f; l_j)$ will inherently have a small value, where l_j is a label of small entropy. Although the characteristics of label entropy values can vary between uniform and skewed cases, it is observed from most real-world multilabel datasets that the skewed case occurs more frequently than uniform case [15]. Additionally, as shown in Figure 2, because MI terms between a feature and labels with small entropy will not much contribute to the final score of the feature, they can be excluded for accelerating multilabel feature selection process.

Figure 3 shows the entropy value of each label in a BibTeX dataset [3] composed of 153 labels; please refer to Table 1 for details. The BibTeX dataset is created from the transactions of user activity in a tag recommendation system. For clarity, we represent the tool which is used to describe and process lists of reference as *BibTeX* whereas the name of the corresponding dataset is BibTeX subsequently. In this system, users freely submit *BibTeX* entries and assign relevant tags. The purpose of this system is recommending a relevant tag for the new *BibTeX* entries submitted by users. The system must identify the relation between *BibTeX* entry and relevant tags based on user transactions previously gathered, and hence, it can be regarded as a real-life text categorization system. For clarity, labels are sorted/ordered according to their entropy value. Figure 3 shows that each label gives a different entropy value; however, more importantly, approximately half of the labels give small entropy values, indicating that the MI terms with those labels will contribute weakly to the final score. Therefore, these labels can be discarded to accelerate the multilabel feature selection process.

Suppose that an algorithm selects $\mathscr{Q} \subset \mathscr{L}$ for reducing computational cost for multilabel feature selection. To prevent possible degradation, i.e., a change in the upper bound for J because of label subset selection, it is preferable that $\mathscr{Q}$ implies a similar upper bound compared to J. In other words, a subset of $\mathscr{L}$ that minimizes

$$\arg\min_{\mathscr{C} \subset \mathscr{L}} \sum_{l \in \mathscr{C}} H(l) = \sum_{l \in \mathscr{L}} H(l) - \sum_{l \in \mathscr{Q}} H(l) \tag{12}$$

is preferable, where $\mathscr{C} = \mathscr{L} \setminus \mathscr{Q}$ is a set of discarded labels.

Proposition 2. *The optimal $\mathscr{C}$ is composed of labels with the lowest entropy.*

Proof 1. Our goal is to identify a subset of labels $\mathscr{C}$ that influences the upper bound of J as insignificantly as possible, when $\mathscr{C}$ is discarded from $\mathscr{L}$ for the feature scoring process. Equation (11) indicates that the upper bound of J is the sum of entropy values for each label and the entropy function always gives positive value, therefore the optimal $\mathscr{L}$ should be composed of labels with the lowest entropy.

Proposition 2 indicates that the optimal $\mathscr{C}$ can be obtained by iteratively discarding a label with the smallest entropy until $\mathscr{Q}$ contains a desirable number of labels. After obtaining $\mathscr{Q}$, the approximated score function for evaluating a feature f is written as

$$\tilde{J}(f) = \sum_{l \in \mathscr{Q}} M(f; l). \tag{13}$$

Finally, the difference between J and $\tilde{J}$ can be exactly calculated as

$$J - \tilde{J} = \sum_{l \in \mathscr{C}} \min(H(f), H(l)), \tag{14}$$

where $J - \tilde{J}$ is always positive because $H(X) \geq 0$. Algorithm 1 describes the procedure of the proposed method.

3.3. Number of Remaining Labels. A final issue related to label subset selection has to do with the number of labels that should be discarded. In fact, because the upper bound of (12) gets larger when the number of discarded labels is increased, there is a trade-off between computational efficiency and the accurate score of each feature. However, the actual computational cost can also be easily predicted after examining some features because the computational cost for examining $|\mathscr{F}|$ features based on (7) is easily calculated as $|\mathscr{F}| \cdot |\mathscr{L}|$, and the computational cost based on (13) is $|\mathscr{F}| \cdot |\mathscr{Q}|$. However, if there is no such constraint and a user only wants to determine a reasonable value of $|\mathscr{Q}|$ for a fast analysis, then a simple and efficient way would be helpful.

Suppose that the algorithm attempts to preserve the upper bound of the score function based on $\mathscr{Q}$, then the upper bound should be greater than or equal to the error because of label subset selection; i.e., the inequality (15) should hold.

$$\sum_{l \in \mathscr{Q}} H(l) \geq \sum_{l \in \mathscr{C}} H(l). \tag{15}$$

Table 1: Standard characteristics of multilabel datasets.

Datasets (abbreviation)	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{F}\rvert$	Feature type	$\lvert L\rvert$	Card	Den	Distinct	PDL	Domain	$\lvert\mathcal{S}\rvert$
BibTeX	7395	1836	Nominal	159	2.402	0.015	2856	0.386	Text	86
Emotions	593	72	Numeric	6	1.868	0.311	27	0.046	Music	24
Enron	1702	1001	Nominal	53	3.378	0.064	753	0.442	Text	41
Genbase	662	1185	Nominal	27	1.252	0.046	32	0.048	Biology	26
Language Log (LLog)	1460	1004	Nominal	75	1.180	0.016	304	0.208	Text	38
Medical	978	1494	Nominal	45	1.245	0.028	94	0.096	Text	31
Slashdot	3782	1079	Nominal	22	1.181	0.054	156	0.041	Text	61
TMC2007	28,596	981	Nominal	22	2.158	0.098	1341	0.047	Text	169
Yeast	2417	103	Numeric	14	4.237	0.303	198	0.082	Biology	49
Arts	7484	1157	Numeric	26	1.654	0.064	599	0.080	Text	87
Business	11,214	1096	Numeric	30	1.599	0.053	233	0.021	Text	106
Computers	12,444	1705	Numeric	33	1.507	0.046	428	0.034	Text	112
Education	12,030	1377	Numeric	33	1.463	0.044	511	0.042	Text	110
Entertainment (entertain)	12,730	1600	Numeric	21	1.414	0.067	337	0.026	Text	113
Health	9205	1530	Numeric	32	1.644	0.051	335	0.036	Text	96
Recreation	12,828	1516	Numeric	22	1.429	0.065	530	0.041	Text	113
Reference	8027	1984	Numeric	33	1.174	0.036	275	0.034	Text	90
Science	6428	1859	Numeric	40	1.450	0.036	457	0.071	Text	80
Social	12,111	2618	Numeric	39	1.279	0.033	361	0.030	Text	110
Society	14,512	1590	Numeric	27	1.670	0.062	1054	0.073	Text	120

Input: n, $|\mathcal{Q}|$;
▷ Number of features to be selected, $n \ll d$
▷ Number of labels to be considered, $|\mathcal{Q}| \ll |\mathcal{L}|$
Output: $\mathcal{S}$; ▷ Selected feature subset, $\mathcal{S}$
Initialize $\mathcal{S} \leftarrow 0$
for all $l \in \mathcal{L}$ **do**
    Calculate value of entropy for l;
end for
Create $\mathcal{Q}$ with $|\mathcal{Q}|$ labels of highest entropy from $\mathcal{L}$;
for all $f \in \mathcal{F}$ **do**
    $\tilde{J}(f) \leftarrow$ Assessing importance of f by using Eq. (13);
end for
Sort $\mathcal{F}$ based upon score values $\tilde{J}$ descendingly;
Set $\mathcal{S} \leftarrow$ Top n features of high score in $\mathcal{F}$;

Algorithm 1: Procedure of Proposed Method.

According to the characteristics of the given labels, the number of labels to be discarded can then be identified as Lemmas 1, 2, and 3.

Lemma 1. *Skewed case.*

$$|\mathcal{Q}| = 1. \tag{16}$$

Proof 2. For simplicity, suppose $\mathcal{L}$ is sorted according to the entropy value of each label, such that l_1 has the smallest entropy and $l_{|\mathcal{L}|}$ has the largest entropy. Suppose that the entropy values of the labels are skewed, as shown in Figure 2. If $l_{|\mathcal{L}|}$ is the only one label with a positive entropy and the remaining labels have no entropy, then the algorithm will move $l_{|\mathcal{L}|}$ to $\mathcal{Q}$ and $l_1, \ldots, l_{|\mathcal{L}|-1}$ to $\mathcal{C}$, and then terminate.

So far, we considered the uniform and skewed cases that are the two extremes of the characteristics in the viewpoint of information content carried by each label. Next, we consider an intermediate between the uniform and skewed cases, in which the information content of each label is proportional to their sequence when they are ascendingly sorted according to their entropy values. For this case, about 30% of labels with the largest entropy should be included in $\mathcal{Q}$.

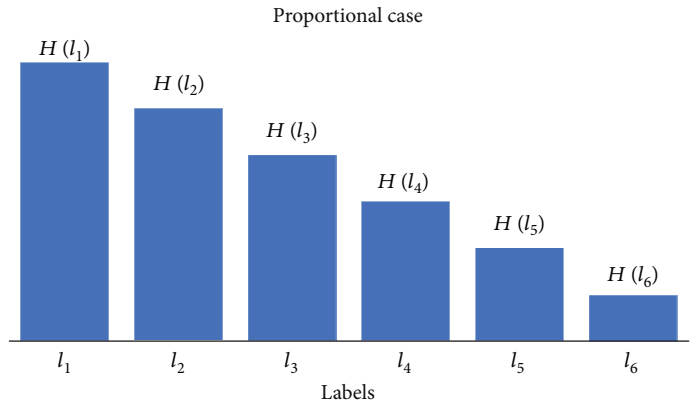


Figure 4: Score value calculation when label entropy values are proportional to their rank.

Lemma 2. *Proportional case.*

$$|\mathcal{Q}| \approx 0.3|\mathcal{L}|. \tag{17}$$

Proof 3. For simplicity, suppose that $\mathcal{L}$ is sorted according to the entropy value of each label, such that l_1 has the smallest entropy value and $l_{|\mathcal{L}|}$ has the largest entropy value. Suppose that the entropy values of the labels are proportional to the sequence number of labels in $\mathcal{L}$ as shown in Figure 4. In this case, an entropy value can be represented as

$$H(l_i) = \alpha \cdot i, \tag{18}$$

where i is the sequence number of label l_i in $\mathcal{L}$. Because the actual entropy value is unnecessary for determining superiority among labels, the term α in (18) can be ignored. Then the entropy value of each label with regard to their sequence can be represented as

$$1, 2, \ldots, |\mathcal{Q}|, \ldots, |\mathcal{L}|. \tag{19}$$

Because the sum of the integers from 1 to i is equal to $i(i+1)/2$, (20) is obtained using (15).

$$\frac{|\mathcal{L}|(|\mathcal{L}|+1)}{2} - \frac{|\mathcal{Q}|(|\mathcal{Q}|+1)}{2} = \frac{|\mathcal{Q}|(|\mathcal{Q}|+1)}{2}. \tag{20}$$

Equation (20) can be simplified as

$$2|\mathcal{Q}|^2 + 2|\mathcal{Q}| - |\mathcal{L}|(|\mathcal{L}|+1) = 0. \tag{21}$$

The solution of (21) is given as

$$\begin{aligned} |\mathcal{Q}| &= \frac{-2 \pm (4 - 4 \cdot 2 \cdot (-|\mathcal{L}|(|\mathcal{L}|+1)))^{1/2}}{4} \\ &= \frac{-2 \pm (4 + 8|\mathcal{L}|(|\mathcal{L}|+1))^{1/2}}{4}. \end{aligned} \tag{22}$$

Because $|\mathcal{Q}|$ is always a positive integer, the negative solution can be ignored. Therefore, we obtain

$$|\mathcal{Q}| = \frac{-2 + (4 + 8|\mathcal{L}|(|\mathcal{L}|+1))^{1/2}}{4}. \tag{23}$$

For clarity, we approximate the solution as

$$\begin{aligned} |\mathcal{Q}| &= \frac{-2 + \left(8|\mathcal{L}|^2 + 8|\mathcal{L}| + 4\right)^{1/2}}{4} \\ &\approx \frac{-2 + \left(2\sqrt{2}|L| + 2\right)^{2\cdot(1/2)}}{4} \\ &= \frac{2\sqrt{2}|L|}{4} \approx 0.7|\mathcal{L}|. \end{aligned} \tag{24}$$

The approximated solution 0.7 $\mathcal{L}$ is slightly greater than the exact solution for $|\mathcal{Q}|$. Therefore, (2) indicates that approximately 70% of labels will be discarded, whereas 30% of labels will remain in $\mathcal{Q}$.

Lemma 3. *Uniform case.*

$$|\mathcal{Q}| = \begin{cases} \dfrac{|\mathcal{L}|}{2}, & \text{if } |\mathcal{L}| \text{ is even,} \\ \dfrac{|\mathcal{L}|}{2+1}, & \text{if } |\mathcal{L}| \text{ is odd.} \end{cases} \tag{25}$$

Proof 4. Suppose that the entropy values of the labels are uniformly distributed as shown in Figure 1. The figure indicates that $|\mathcal{Q}|$ should have corresponding labels with regard to each discarded label. Therefore, for the even case, the number of labels in $\mathcal{Q}$ and $\mathcal{C}$ must be the same for (15) to hold; thus, $|\mathcal{Q}| = |\mathcal{L}|/2$. For the odd case, $\mathcal{Q}$ must have one more label than $\mathcal{C}$; thus, $|\mathcal{Q}| = |\mathcal{L}|/2 + 1$.

The proof indicates that the number of labels to be selected is decreased as the entropy values of labels are skewed. In addition, the proof guarantees that $|\mathcal{Q}|$ must be lesser than $|\mathcal{L}|$ and the computational cost for evaluating

the importance of each feature based on $\mathcal{Q}$ must be smaller than $|L|/2 + 1$. Therefore, Theorem 1 can be obtained.

Theorem 1 *$|\mathcal{Q}|$ is always smaller than $|\mathcal{L}|$.*

Proof 5. Suppose that there are two label sets $\mathcal{Q}$ and $\mathcal{C}$ to be considered and ignored for calculating the importance of each feature, respectively. Because $\mathcal{Q}$ should carry the majority information content than $\mathcal{C}$, $\sum_{l\in\mathcal{Q}} H(l)$ should be larger than $\sum_{l\in\mathcal{C}} H(l)$. As shown in Proposition 2, the algorithm is able to achieve this goal by (1) including a label with the largest entropy in $\mathcal{Q}$ and removing that label from $\mathcal{L}$, (2) including labels with the smallest entropy in $\mathcal{C}$ and removing those labels from $\mathcal{L}$ iteratively until $\sum_{l\in\mathcal{Q}} H(l) > \sum_{l\in\mathcal{C}} H(l)$, and (3) repeating (1) to (2) until $\mathcal{L}$ has no element. If the entropy values of all the labels are the same, i.e., the largest entropy value and the smallest entropy value are the same, one label can be included in $\mathcal{C}$ when a label is included in $\mathcal{Q}$ as Lemma 3. Thus, $\mathcal{C}$ possibly has more labels than $\mathcal{Q}$ in the case when the smallest entropy value is actually smaller than the largest entropy value, indicating that the uniform case is the worst case from the viewpoint of the number of labels in $\mathcal{Q}$. Consequently, the number of labels in $|\mathcal{Q}|$ cannot be larger than $|\mathcal{L}|/2 + 1$.

Because $|\mathcal{Q}|$ is always smaller than $|\mathcal{L}|$ and calculating one MI term is regarded as the unit cost, the computational cost for evaluating each feature using $\tilde{J}$ is constant in the viewpoint of the number of labels.

3.4. Influence to Feature Ranking. The multilabel feature selection is done by ranking each feature according to its importance value. After label subset selection is conducted, the importance score of each feature will be calculated by summing $M(f; l_i)$ terms, where $l_i \in \mathcal{Q}$. However, when the entropy values of labels are skewed, the rank based on J and that based on $\tilde{J}$ are unlikely to change. To demonstrate this aspect, we illustrate how the importance score is calculated under the skewed case in Figure 5. In the figure, there are three labels, namely l_1, l_2, and l_3; l_1 has the highest entropy, whereas l_2 and l_3 have insignificant entropies. The MI between each feature and each label is represented as yellow bars, and the final score of each feature is represented on the right hand side of the figure. The figure indicates that (1) the MI between each feature and each label is bound by the entropy of each label and (2) the MI between each feature and the labels of high entropy mostly determines the final score of each feature. In other words, (3) the influence of MIs between each feature and l_2 and l_3 is insignificant to the final score.

With regard to the process of feature selection, Figure 5 implies three more indications. The first indication is related to the influence of labels with high entropy to the final score. Because the final score is determined by summing MI terms between a feature and all the labels, a feature that is dependent on labels with high entropy is likely to have a high importance score. Therefore, those features will be included to the final feature subset $\mathcal{S}$

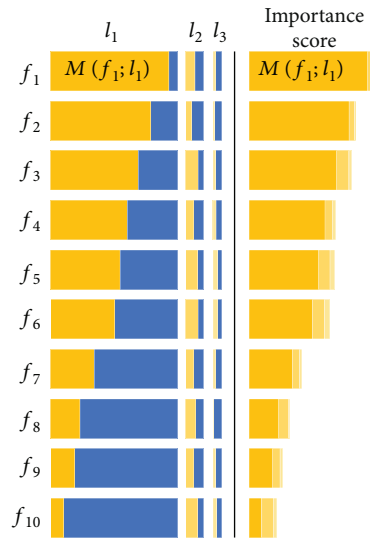


Figure 5: Importance score of each feature in the viewpoint of entropy of each label when entropy values of labels are skewed.

Figure 6: An example of rank change.

because of their higher rank, and they show promise as potential members of $\mathcal{S}$. The second indication is related to the change among similarly ranked features. However, because the goal of feature selection is to select a feature subset that is composed of n features, the specific rank of each feature is unimportant. For example, suppose that the algorithm tries to choose ten features from $\mathcal{S}$ because $\mathcal{F}$ is set to ten by users or there is a limitation on the storage. The label subset selection may change the rank of the second- and the third-ranked features; however, these two features will be included in the final feature subset $\mathcal{S}$ because the algorithm is allowed to select ten features. The final indication is related to the rank among unimportant features. Although there may be a set of features that are dependent on labels with small entropy, these features will have low importance scores and hence will be discarded from $\mathcal{S}$.

Although the example of Figure 6 indicates that the rank of each feature will be unlikely to change or may be changed meaninglessly, empirical experiments should be followed to investigate the availability of label subset selection.

4. Experimental Results

A description of the multilabel datasets, algorithms, statistical tests, and other settings used in the experimental study is provided in this section. Next, the experimental results based on different multilabel learning methods, the datasets, and the analysis are presented subsequently.

4.1. Experimental Settings. Twenty real multilabel datasets were employed in our experiments [12, 25, 35], where the number of relevant and irrelevant features is unknown. Table 1 shows the standard statistics of the multilabel datasets and the meaning of each notation is described as follows:

(i) $|\mathcal{U}|$: number of patterns in the dataset

(ii) $|\mathcal{F}|$: number of features

(iii) Feature type: type of feature

(iv) $|\mathcal{L}|$: number of labels

(v) Card: average number of labels for each instance (label cardinality)

(vi) Den: label cardinality divided by the total number of labels (label density)

(vii) Distinct: number of unique label subsets in $\mathcal{L}$ (distinct label set)

(viii) *PDL*: number of distinct label sets divided by the total number of patterns (portion of distinct labels)

(ix) Domain: applications to which each dataset corresponds

(x) $|\mathcal{S}|$: number of features to be selected ($\sqrt{|\mathcal{W}|}$)

These statistics show that the 20 datasets cover a broad range of cases with diversified multilabel properties. In the case where the feature type is numeric, we discretized the features using the LAIM discretization method [45]. In addition, datasets that are composed of more than 10,000 features are preprocessed to contain the top 2% and 5% features with the highest document frequency [12, 46]. We conducted an 8 : 2 hold-out cross-validation, and each experiment was repeated ten times. The average value was taken to represent the classification performance. A wide variety of multilabel classifiers can be considered to conduct multilabel classification [8]. In this study, we chose the multilabel naive Bayes classifier [41] because the learning process can be conducted quickly, owing to the well-known naive Bayes assumption, without incurring an additional tuning process, and because our primary concern in this study is efficient multilabel learning. Finally, we considered four evaluation measures, which are employed in many multilabel learning studies: execution time for the training and test process, Hamming loss, multilabel accuracy, and subset accuracy [8, 29].

The Friedman test was employed to analyze the performance of the multilabel feature selection methods; it is a widely used statistical test for comparing multiple methods over a number of datasets [47]. The null hypothesis of the equal performance of the compared algorithms is rejected in terms of each evaluation measure if the Friedman statistic F_F is greater than the critical value at significance level α. In this case, we need to proceed with certain post hoc tests to analyze the relative performance of the comparison methods. The Bonferroni-Dunn test is employed because we are interested in determining whether the proposed method achieves a performance similar to that of the feature selection process considering all of the labels and to that of the multilabel learning without the feature selection process [48]. For the Bonferroni-Dunn test, the performances of the proposed method and another method are deemed to be statistically similar if their average ranks over all datasets are within one CD. For our experiments, the critical value at the significance level $\alpha = 0.05$ is 2.492, and the CD with $\alpha = 0.05$ is 1.249 because $q_{0.05} = 2.498$ [48].

4.2. Comparative Studies. In this section, we compare the proposed feature selection method based on the label subset selection strategy to the conventional multilabel learning without the feature selection process and the conventional feature selection method without the label subset selection. The detail of each method, besides the proposed method, is described as follows:

(i) No: conventional multilabel learning the without feature selection process. Here, $\mathcal{F}$ is used as the input features for the multilabel classifier

(ii) SL: multilabel learning with the proposed feature selection process. Here, S is used as the input features. In the feature selection process, only one label with the highest entropy is considered to measure the importance of each feature

(iii) 3L: multilabel learning with the proposed feature selection process. Here, S is used as the input features. In the feature selection process, 30% of labels with the highest entropy are chosen by the label selection strategy to compose Q

(iv) 5L: multilabel learning with the proposed feature selection process. Here, S is used as the input features. In the feature selection process, 50% of labels with the highest entropy are chosen by the label selection strategy to compose Q

(v) AL: multilabel learning with the conventional feature selection process. Here, S is used as the input features. The same feature subset can be obtained by setting $Q = L$ for the proposed method

All methods were carefully implemented in a MATLAB 8.2 programming environment and tested on an Intel Core i7-3930 K (3.2 GHz) with 64 GB memory.

Tables 2–5 report the detailed experimental results of each method under comparison on 20 multilabel datasets. For each evaluation measure, ↓ means *the smaller the better* whereas ↑ means *the larger the better*. The best

Table 2: Execution time (↓) for training and testing process of each comparing method (mean ± std. deviation) on 20 multilabel datasets.

Method	BibTeX	Emotions	Enron	Genbase	LLog	Medical	Slashdot
No	141.852 ± 0.386	0.091 ± 0.001	12.819 ± 0.028	4.713 ± 0.032	17.053 ± 0.057	12.996 ± 0.090	7.796 ± 0.020
SL	**9.326 ± 0.322** ·	0.069 ± 0.050	**1.039 ± 0.198** ·	**0.541 ± 0.233** ·	**1.164 ± 0.201** ·	**0.870 ± 0.281** ·	**1.279 ± 0.209** ·
3L	17.820 ± 1.838	0.058 ± 0.018	1.846 ± 0.592	0.980 ± 0.560	2.194 ± 0.719	1.846 ± 0.925	2.241 ± 0.502
5L	20.686 ± 2.355	0.070 ± 0.028	2.118 ± 0.734	1.206 ± 0.740	2.455 ± 0.859	2.176 ± 1.149	2.239 ± 0.503
AL	201.458 ± 41.405	**0.038 ± 0.007** ·	3.112 ± 1.768	1.071 ± 0.903	4.027 ± 2.622	3.353 ± 2.561	3.742 ± 1.166
Method	TMC2007	Yeast	Arts	Business	Computers	Education	Entertain
No	28.134 ± 0.033	0.450 ± 0.004	15.814 ± 0.094	20.921 ± 0.166	38.178 ± 0.369	29.887 ± 0.320	23.983 ± 0.324
SL	**8.962 ± 0.258** ·	0.291 ± 0.056	**2.750 ± 0.182** ·	**4.158 ± 0.291** ·	**6.140 ± 0.318** ·	**5.093 ± 0.350** ·	**5.269 ± 0.280** ·
3L	13.903 ± 0.662	0.314 ± 0.039	4.697 ± 0.508	6.755 ± 0.609	10.701 ± 0.805	8.440 ± 0.910	9.568 ± 0.729
5L	13.908 ± 0.665	0.347 ± 0.054	5.688 ± 0.664	8.047 ± 0.783	12.995 ± 1.066	10.115 ± 1.213	9.572 ± 0.724
AL	84.744 ± 7.584	**0.238 ± 0.012** ·	15.074 ± 2.802	32.428 ± 8.005	97.128 ± 12.655	55.539 ± 10.467	63.178 ± 6.986
Method	Health	Recreation	Reference	Science	Social	Society	Avg. rank
No	28.702 ± 0.206	24.032 ± 0.025	34.620 ± 0.293	34.302 ± 0.050	67.920 ± 0.128	34.704 ± 0.032	4.40
SL	**4.258 ± 0.267** ·	**5.165 ± 0.286** ·	**4.246 ± 0.354** ·	**3.828 ± 0.414** ·	**8.281 ± 0.691** ·	**6.655 ± 0.337** ·	**1.15** ·
3L	7.272 ± 0.689	9.310 ± 0.721	7.495 ± 0.885	6.709 ± 1.133	15.099 ± 1.944	11.544 ± 0.820	2.10
5L	8.780 ± 0.915	9.311 ± 0.720	9.137 ± 1.168	8.154 ± 1.507	18.516 ± 2.596	13.989 ± 1.075	3.10
AL	44.589 ± 6.295	59.392 ± 7.450	54.074 ± 11.295	49.855 ± 12.972	263.023 ± 51.640	107.083 ± 14.488	4.25

Table 3: Hamming loss (↓) performance of each comparing method (mean ± std. deviation) on 20 multilabel datasets.

Method	BibTeX	Emotions	Enron	Genbase	LLog	Medical	Slashdot
No	0.082 ± 0.002	**0.240 ± 0.028** ·	0.214 ± 0.009	**0.007 ± 0.001** ·	0.340 ± 0.024	0.019 ± 0.001	**0.041 ± 0.001** ·
SL	**0.067 ± 0.002** ·	0.268 ± 0.020	0.144 ± 0.005	0.008 ± 0.001	**0.201 ± 0.013** ·	0.032 ± 0.003	0.047 ± 0.002
3L	0.071 ± 0.003	0.266 ± 0.023	**0.139 ± 0.005** ·	0.007 ± 0.001	0.250 ± 0.010	**0.014 ± 0.002** ·	0.044 ± 0.001
5L	0.080 ± 0.002	0.266 ± 0.025	0.140 ± 0.004	0.008 ± 0.002	0.254 ± 0.011	0.015 ± 0.002	0.043 ± 0.002
AL	0.086 ± 0.001	0.265 ± 0.023	0.140 ± 0.003	0.010 ± 0.003	0.253 ± 0.010	0.018 ± 0.002	0.043 ± 0.002
Method	TMC2007	Yeast	Arts	Business	Computers	Education	Entertain
No	0.139 ± 0.001	0.272 ± 0.007	0.109 ± 0.004	0.090 ± 0.002	0.117 ± 0.003	0.079 ± 0.002	0.123 ± 0.004
SL	**0.107 ± 0.001** ·	0.271 ± 0.007	0.072 ± 0.002	**0.050 ± 0.002** ·	0.080 ± 0.003	**0.055 ± 0.002** ·	0.111 ± 0.003
3L	0.125 ± 0.002	**0.270 ± 0.005** ·	0.072 ± 0.002	0.067 ± 0.002	**0.064 ± 0.003** ·	0.058 ± 0.002	**0.078 ± 0.002** ·
5L	0.126 ± 0.001	0.273 ± 0.007	**0.071 ± 0.002** ·	0.069 ± 0.003	0.068 ± 0.003	0.058 ± 0.002	0.081 ± 0.002
AL	0.123 ± 0.001	0.276 ± 0.007	0.072 ± 0.002	0.070 ± 0.003	0.070 ± 0.003	0.059 ± 0.002	0.081 ± 0.002
Method	Health	Recreation	Reference	Science	Social	Society	Avg. rank
No	0.073 ± 0.002	0.129 ± 0.005	0.097 ± 0.003	0.132 ± 0.004	0.077 ± 0.002	0.197 ± 0.003	4.20
SL	0.055 ± 0.002	**0.063 ± 0.001** ·	0.079 ± 0.004	0.056 ± 0.003	**0.040 ± 0.002** ·	0.173 ± 0.005	2.80
3L	0.056 ± 0.001	0.073 ± 0.002	**0.066 ± 0.003** ·	**0.054 ± 0.004** ·	0.045 ± 0.002	0.144 ± 0.007	**2.25** ·
5L	**0.053 ± 0.001** ·	0.071 ± 0.002	0.070 ± 0.004	0.055 ± 0.003	0.051 ± 0.002	0.135 ± 0.007	2.60
AL	0.053 ± 0.002	0.073 ± 0.003	0.071 ± 0.004	0.057 ± 0.003	0.052 ± 0.002	**0.134 ± 0.007** ·	3.15

performance among the five methods under comparison is shown in boldface with a bullet mark. In addition, the average rank of each method under comparison over all the multilabel datasets is presented in the last column of each table. Table 6 represents the Friedman statistics F_F and the corresponding critical values on each evaluation measure. As shown in Table 6, at significance level $\alpha = 0.05$, the null hypothesis of *equal* performance among the methods under comparison is clearly rejected in terms of each evaluation measure.

To show the relative performance of the proposed method and conventional multilabel learning methods,

Table 4: Multilabel accuracy (↑) performance of each comparing method (mean ± std. deviation) on 20 multilabel datasets.

Method	BibTeX	Emotions	Enron	Genbase	LLog	Medical	Slashdot
No	**0.191 ± 0.006** ·	**0.543 ± 0.043** ·	0.196 ± 0.008	0.904 ± 0.019	0.037 ± 0.001	0.335 ± 0.029	**0.445 ± 0.014** ·
SL	0.115 ± 0.006	0.486 ± 0.030	0.229 ± 0.011	0.917 ± 0.018	**0.053 ± 0.004** ·	0.517 ± 0.041	0.265 ± 0.019
3L	0.171 ± 0.008	0.488 ± 0.036	**0.236 ± 0.009** ·	**0.924 ± 0.019** ·	0.044 ± 0.002	**0.705 ± 0.029** ·	0.345 ± 0.012
5L	0.166 ± 0.007	0.490 ± 0.037	0.235 ± 0.009	0.919 ± 0.017	0.043 ± 0.002	0.690 ± 0.030	0.364 ± 0.014
AL	0.162 ± 0.008	0.489 ± 0.036	0.235 ± 0.008	0.919 ± 0.019	0.043 ± 0.002	0.667 ± 0.042	0.362 ± 0.014
Method	TMC2007	Yeast	Arts	Business	Computers	Education	Entertain
No	0.395 ± 0.004	**0.425 ± 0.010** ·	**0.328 ± 0.007** ·	0.627 ± 0.006	0.338 ± 0.006	**0.319 ± 0.008** ·	0.348 ± 0.008
SL	0.410 ± 0.005	0.414 ± 0.011	0.225 ± 0.018	**0.666 ± 0.009** ·	0.399 ± 0.013	0.233 ± 0.008	0.294 ± 0.004
3L	0.417 ± 0.005	0.422 ± 0.010	0.281 ± 0.011	0.649 ± 0.008	**0.434 ± 0.007** ·	0.267 ± 0.008	**0.405 ± 0.004** ·
5L	0.416 ± 0.004	0.419 ± 0.010	0.296 ± 0.009	0.648 ± 0.007	0.434 ± 0.008	0.269 ± 0.009	0.391 ± 0.009
AL	**0.430 ± 0.004** ·	0.416 ± 0.009	0.300 ± 0.011	0.644 ± 0.008	0.431 ± 0.009	0.268 ± 0.007	0.393 ± 0.010
Method	Health	Recreation	Reference	Science	Social	Society	Avg. rank
No	0.476 ± 0.006	0.343 ± 0.011	0.388 ± 0.020	0.215 ± 0.006	0.516 ± 0.009	**0.202 ± 0.004** ·	3.40
SL	0.514 ± 0.004	0.294 ± 0.005	0.410 ± 0.009	0.163 ± 0.016	0.480 ± 0.009	0.168 ± 0.005	4.25
3L	0.516 ± 0.008	0.352 ± 0.018	**0.432 ± 0.006** ·	0.223 ± 0.016	0.542 ± 0.011	0.185 ± 0.007	2.40
5L	**0.518 ± 0.004** ·	**0.369 ± 0.006** ·	0.432 ± 0.008	**0.229 ± 0.010** ·	**0.544 ± 0.009** ·	0.191 ± 0.006	**2.25** ·
AL	0.516 ± 0.003	0.362 ± 0.010	0.431 ± 0.008	0.223 ± 0.014	0.544 ± 0.009	0.192 ± 0.006	2.70

Table 5: Subset accuracy (↑) performance of each comparing method (mean ± std. deviation) on 20 multilabel datasets.

Method	BibTeX	Emotions	Enron	Genbase	LLog	Medical	Slashdot
No	0.063 ± 0.005	**0.242 ± 0.049** ·	0.001 ± 0.001	**0.863 ± 0.027** ·	0.000 ± 0.000	0.301 ± 0.027	**0.357 ± 0.016** ·
SL	0.048 ± 0.006	0.181 ± 0.041	0.003 ± 0.003	0.833 ± 0.032	**0.002 ± 0.002** ·	0.319 ± 0.042	0.233 ± 0.017
3L	0.062 ± 0.006	0.186 ± 0.031	0.004 ± 0.004	0.842 ± 0.034	0.000 ± 0.000	**0.551 ± 0.041** ·	0.298 ± 0.015
5L	0.063 ± 0.006	0.181 ± 0.035	**0.005 ± 0.005** ·	0.835 ± 0.030	0.000 ± 0.000	0.531 ± 0.038	0.311 ± 0.014
AL	**0.064 ± 0.006** ·	0.181 ± 0.037	**0.005 ± 0.005** ·	0.835 ± 0.033	0.000 ± 0.000	0.510 ± 0.052	0.311 ± 0.015
Method	TMC2007	Yeast	Arts	Business	Computers	Education	Entertain
No	0.086 ± 0.005	0.098 ± 0.007	0.164 ± 0.008	0.469 ± 0.014	0.138 ± 0.007	0.179 ± 0.008	0.171 ± 0.008
SL	**0.119 ± 0.003** ·	0.093 ± 0.009	0.146 ± 0.018	**0.504 ± 0.013** ·	0.275 ± 0.019	0.176 ± 0.007	0.150 ± 0.004
3L	0.106 ± 0.005	**0.098 ± 0.010** ·	0.195 ± 0.011	0.490 ± 0.011	**0.335 ± 0.008** ·	0.192 ± 0.007	**0.283 ± 0.012** ·
5L	0.107 ± 0.003	0.096 ± 0.009	0.203 ± 0.010	0.489 ± 0.011	0.332 ± 0.009	**0.193 ± 0.007** ·	0.250 ± 0.020
AL	0.115 ± 0.004	0.093 ± 0.008	**0.206 ± 0.012** ·	0.486 ± 0.012	0.328 ± 0.010	0.191 ± 0.006	0.249 ± 0.020
Method	Health	Recreation	Reference	Science	Social	Society	Avg. rank
No	0.227 ± 0.008	0.140 ± 0.009	0.240 ± 0.035	0.072 ± 0.006	0.402 ± 0.014	**0.069 ± 0.003** ·	3.68
SL	**0.336 ± 0.008** ·	0.223 ± 0.004	0.355 ± 0.009	0.104 ± 0.016	0.389 ± 0.013	0.038 ± 0.006	3.78
3L	0.329 ± 0.009	0.269 ± 0.016	0.375 ± 0.006	0.148 ± 0.009	0.456 ± 0.013	0.055 ± 0.008	2.63
5L	0.336 ± 0.006	**0.285 ± 0.008** ·	**0.376 ± 0.008** ·	**0.158 ± 0.008** ·	**0.460 ± 0.012** ·	0.055 ± 0.008	**2.23** ·
AL	0.333 ± 0.006	0.284 ± 0.010	0.374 ± 0.008	0.151 ± 0.010	0.456 ± 0.011	0.055 ± 0.007	2.70

Figure 7 illustrates the CD diagrams on each evaluation measure, where the average rank of each method is marked along the axis with better ranks placed on the right hand side of each figure [47]. In each figure, any comparison method whose average rank is within one CD to that of the best method is interconnected with a thick line; the length of the thick line indicates the extent of CD on a diagram. Otherwise, any method not connected with the best method is considered to have a significantly different performance from the latter.

Based on the empirical experiments and statistical analysis, the following indications can be observed:

Table 6: Summary of the Friedman statistics F_F ($k = 5$, $N = 20$) and the critical value in terms of each evaluation measure.

Evaluation measure	F_F	Critical value ($\alpha = 0.05$)
Execution time	66.011	2.492
Hamming loss	5.437	
Multilabel accuracy	7.153	
Subset accuracy	4.421	

(1) As Figure 7 shows, the multilabel learning and classification process is significantly accelerated by the feature selection process. In particular, the multilabel classification with SL and 3L is completed significantly faster than No, indicating the superiority of the proposed approach

(2) Focusing on the average rank of AL and No in Figure 7, the advantage of multilabel feature selection from the viewpoint of the execution time is insignificant, indicating that the merit given by feature selection process on the execution time can disappear owing to a large number of labels

(3) As Figure 7 shows, the feature subset selected by AL is able to deliver a statistically similar classification performance to the baseline performance No. This means that the dimensionality of the input space can be reduced to accelerate the multilabel learning process without degrading the predictive performance

(4) The feature subset selected by the proposed methods based on the label subset selection such as 3L and 5L is able to deliver a comparable classification performance to the classifier if a moderate number of labels are considered for evaluating the importance of features

(5) A notable exception can be observed from the experimental results of SL, which considers only one label for the feature scoring process. However, it also gives a statistically better performance than No in the experiments involving Hamming loss and a comparable performance in the experiments involving multilabel accuracy and subset accuracy

(6) Surprisingly, if a moderate number of labels are considered from the feature scoring process like 3L or 5L, the feature subset gives statistically better discriminating power than the baseline performance given by No. For example, in the experiments involving Hamming loss, as shown in Table 3, 3L gives a better Hamming loss performance than No on 85% of multilabel datasets

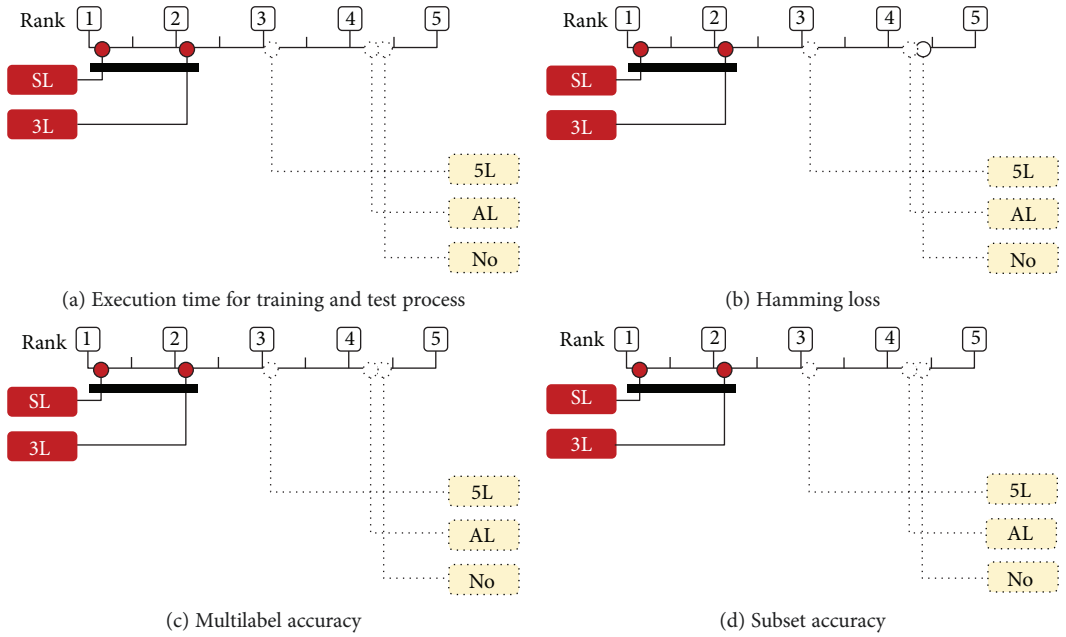
(7) Furthermore, based on the comparison to the multilabel classification performance given by No, the feature subset selected by 3L gives a better Hamming loss performance on 70% of multilabel datasets. This tendency can be observed again from the experiments involving multilabel accuracy based on 5L as it gives a better performance on 80% of datasets

In summary, the experimental results show that the proposed method based on the label subset selection strategy achieves a significantly better execution time than the baseline multilabel setting No and conventional multilabel learning with feature selection AL, indicating that the proposed method is able to accelerate the multilabel learning process. Furthermore, the feature subset selected by the proposed method, such as 3L and 5L, yields a similar classification performance compared to the other methods. Because the proposed method has a lower execution time compared to the other methods, this means that the proposed method is able to quickly identify the important feature subset, without degrading the multilabel classification performance.

Finally, we conducted additional experiments to validate the scalability and efficiency of the proposed method. For this purpose, we employed the Delicious dataset, which is composed of a large number of patterns and labels [3]. Specifically, the Delicious dataset was extracted from the del.icio.us social bookmarking site where textual patterns and associated labels represent web pages and relevant tags. This dataset is composed of 16,105 patterns, 500 features, and 983 labels from 15,806 unique label subsets. To demonstrate the superiority of the proposed method, we employed MLCFS [19] and PPT + RF [24]. In this experiment, we regard 3L as the proposed method because it performs better than SL, 5L, and AL, as shown in Figure 7. Table 7 represents the experimental results of three multilabel feature selection methods, including the proposed method. The experimental results indicate that the proposed method outputs the final feature subset much faster than the compared methods with similar multilabel classification performances in terms of Hamming loss, multilabel accuracy, and subset accuracy.

5. Conclusion

In this study, we proposed an efficient multilabel feature selection method to achieve scalable multilabel learning when the number of labels is large. Because the computational load of the multilabel learning process increases with the increasing number of features in the input data, the proposed method accelerates the multilabel learning process by selecting important features to reduce the dimensionality of features. In addition, with regard to the multiple labels considered for the feature scoring process, we demonstrated that the feature selection process itself can be accelerated for further acceleration of the multilabel learning process. Furthermore, empirical experiments on 20 multilabel datasets showed that the multilabel learning process can be boosted without deteriorating the discriminating power of the multilabel classifier.



(a) Execution time for training and test process

(b) Hamming loss

(c) Multilabel accuracy

(d) Subset accuracy

Figure 7: Bonferroni-Dunn test results of five comparing methods with four evaluation measures. Methods not connected with the best method in the CD diagram are considered to have significantly different performance (significance level $\alpha = 0.05$). This is reproduced from Lee et al. (2017) (under the Creative Commons Attribution License/public domain).

Table 7: Comparison results of proposed method, MLCFS, and PPT + RF on the Delicious dataset.

Methods	Execution time	Hamming loss	Multilabel accuracy	Subset accuracy
Proposed method (3L)	**26.6326 ± 0.9547** ·	0.0201 ± 0.0002	0.0301 ± 0.0002	0.0001 ± 0.0001
MLCFS	144.0414 ± 13.3807	0.0201 ± 0.0002	0.0304 ± 0.0043	0.0001 ± 0.0002
PPT + RF	1556.1397 ± 30.1202	0.0201 ± 0.0002	0.0301 ± 0.0054	0.0002 ± 0.0003

Future research directions include scalability against a large number of training examples. Although this can be achieved by a multilabel classification approach using distributed computing [49], the performance should be tested empirically to validate the potential. In addition, we will investigate the multilabel learning performance with respect to the label selection strategy. Our experiments indicate that the feature subset selected by the proposed method can possibly deliver a better discriminating capability, despite only a part of the labels in a given label set being considered for the feature scoring process. Because this was an unexpected result, as the primary goal of this study was the acceleration of the multilabel learning process, we would like to investigate this issue more thoroughly in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

Both authors declare that they have no conflicts of interest.

Acknowledgments

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

References

[1] J. Paulin, A. Calinescu, and M. Wooldridge, "Agent-based modeling for complex financial systems," *IEEE Intelligent Systems*, vol. 33, no. 2, pp. 74–82, 2018.

[2] G. Le Moal, G. Moraru, P. Véron, P. Rabaté, and M. Douilly, "Feature selection for complex systems monitoring: an application using data fusion," in *CCCA12*, pp. 1–6, Marseilles, France, December 2012.

[3] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels," in *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data*, pp. 30–44, Antwerp, Belgium, 2008.

[4] M. Zanin, E. Menasalvas, S. Boccaletti, and P. Sousa, "Feature selection in the reconstruction of complex network representations of spectral data," *PLoS ONE*, vol. 8, no. 8, p. e72045, 2013.

[5] J. Lee, J. Chae, and D.-W. Kim, "Effective music searching approach based on tag combination by exploiting prototypical

acoustic content," *Multimedia Tools and Applications*, vol. 76, no. 4, pp. 6065–6077, 2017.

[6] T. Pfingsten, D. J. L. Herrmann, T. Schnitzler, A. Feustel, and B. Scholkopf, "Feature selection for troubleshooting in complex assembly lines," *IEEE Transactions on Automation Science and Engineering*, vol. 4, no. 3, pp. 465–469, 2007.

[7] T. Rault, A. Bouabdallah, Y. Challal, and F. Marin, "A survey of energy-efficient context recognition systems using wearable sensors for healthcare applications," *Pervasive and Mobile Computing*, vol. 37, pp. 23–44, 2017.

[8] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[9] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang, "Document transformation for multi-label feature selection in text categorization," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 451–456, Omaha, NE, USA, October 2007.

[10] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in Neural Information Processing Systems 14*, pp. 681–687, Vancouver, Canada, 2001.

[11] Y. Yu, W. Pedrycz, and D. Miao, "Multi-label classification by exploiting label correlations," *Expert Systems with Applications*, vol. 41, no. 6, pp. 2989–3004, 2014.

[12] M.-L. Zhang and L. Wu, "Lift: multi-label learning with label-specific features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 107–120, 2015.

[13] F. Li, D. Miao, and W. Pedrycz, "Granular multi-label feature selection based on mutual information," *Pattern Recognition*, vol. 67, pp. 410–423, 2017.

[14] J. Lee and D.-W. Kim, "SCLS: multi-label feature selection based on scalable criterion for large label set," *Pattern Recognition*, vol. 66, pp. 342–352, 2017.

[15] J. Lee and D.-W. Kim, "Fast multi-label feature selection based on information-theoretic feature ranking," *Pattern Recognition*, vol. 48, no. 9, pp. 2761–2771, 2015.

[16] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1079–1089, 2011.

[17] N. Spolaôr, M. C. Monard, G. Tsoumakas, and H. D. Lee, "A systematic review of multi-label feature selection and a new method based on label construction," *Neurocomputing*, vol. 180, no. 1, pp. 3–15, 2016.

[18] N. Spolaôr, E. A. Cherman, M. C. Monard, and H. D. Lee, "A comparison of multi-label feature selection methods using the problem transformation approach," *Electronic Notes in Theoretical Computer Science*, vol. 292, pp. 135–151, 2013.

[19] S. Jungjit, M. Michaelis, A. A. Freitas, and J. Cinatl, "Two extensions to multi-label correlation-based feature selection: a case study in bioinformatics," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1519–1524, Manchester, UK, October 2013.

[20] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with naïve Bayes," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 5432–5435, 2009.

[21] J. Read, "A pruned problem transformation method for multi-label classification," in *Proc. 2008 New Zealand Computer Science Research Student Conference*, pp. 143–150, Christchurch, New Zealand, April 2008.

[22] G. Doquire and M. Verleysen, "Mutual information-based feature selection for multilabel classification," *Neurocomputing*, vol. 122, pp. 148–155, 2013.

[23] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of Relief F and RReliefF," *Machine Learning*, vol. 53, no. 1/2, pp. 23–69, 2003.

[24] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann, "Categorizing feature selection methods for multi-label classification," *Artificial Intelligence Review*, vol. 49, no. 1, pp. 57–78, 2018.

[25] O. Reyes, C. Morell, and S. Ventura, "Scalable extensions of the relieff algorithm for weighting and selecting features on the multi-label learning context," *Neurocomputing*, vol. 161, pp. 168–182, 2015.

[26] Y. Sun, A. Wong, and M. S. Kamel, "Classification of imbalanced data: a review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009.

[27] Q. Gu, Z. Li, and J. Han, "Correlated multi-label feature selection," in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, pp. 1087–1096, Glasgow, Scotland, UK, October 2011.

[28] X. Kong and P. S. Yu, "gMLC: a multi-label feature selection framework for graph classification," *Knowledge and Information Systems*, vol. 31, no. 2, pp. 281–305, 2012.

[29] J. Lee and D.-W. Kim, "Feature selection for multi-label classification using multivariate mutual information," *Pattern Recognition Letters*, vol. 34, no. 3, pp. 349–357, 2013.

[30] 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.

[31] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint ℓ2,1-norms minimization," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1813–1821, 2010.

[32] S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in *Proc. 21th Int. Joint Conf. Artificial Intelligence*, pp. 1077–1082, Pasadena, USA, July 2009.

[33] B. Qian and I. Davidson, "Semi-supervised dimension reduction for multi-label classification," in *Proc. 24th AAAI Conf. Artificial Intelligence*, pp. 569–574, Atlanta, USA, July 2010.

[34] D. Kong, C. Ding, H. Huang, and H. Zhao, "Multi-label reliefF and F-statistic feature selections for image annotation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2352–2359, Providence, RI, USA, June 2012.

[35] J. Lee and D.-W. Kim, "Memetic feature selection algorithm for multi-label classification," *Information Sciences*, vol. 293, pp. 80–96, 2015.

[36] J. Lee and D.-W. Kim, "Mutual information-based multi-label feature selection using interaction information," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2013–2025, 2015.

[37] J. Lee, H. Lim, and D.-W. Kim, "Approximating mutual information for multi-label feature selection," *Electronics Letters*, vol. 48, no. 15, pp. 929-930, 2012.

[38] H. Lim, J. Lee, and D.-W. Kim, "Multi-label learning using mathematical programming," *IEICE Transactions on Information and Systems*, vol. E98.D, no. 1, pp. 197–200, 2015.

[39] J. Lee, W. Seo, and D. W. Kim, "Effective evolutionary multilabel feature selection under a budget constraint," *Complexity*, vol. 2018, Article ID 3241489, 14 pages, 2018.

[40] J. Lee and D.-W. Kim, "Efficient multi-label feature selection using entropy-based label selection," *Entropy*, vol. 18, no. 11, p. 405, 2016.

[41] M.-L. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive bayes classification," *Information Sciences*, vol. 179, no. 19, pp. 3218–3229, 2009.

[42] I. Pillai, G. Fumera, and F. Roli, "Designing multi-label classifiers that maximize f measures: state of the art," *Pattern Recognition*, vol. 61, pp. 394–404, 2017.

[43] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[44] T. Cover and J. Thomas, *Elements of Information Theory*, vol. 6, Wiley Online Library, New York, 1991.

[45] A. Cano, J. M. Luna, E. L. Gibaja, and S. Ventura, "Laim discretization for multi-label data," *Information Sciences*, vol. 330, pp. 370–384, 2016.

[46] Y. Yang and S. Gopal, "Multilabel classification with meta-level features in a learning-to-rank framework," *Machine Learning*, vol. 88, no. 1-2, pp. 47–68, 2012.

[47] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 1–30, 2006.

[48] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.

[49] J. Gonzalez-Lopez, S. Ventura, and A. Cano, "Distributed nearest neighbor classification for large-scale multi-label data on spark," *Future Generation Computer Systems*, vol. 87, pp. 66–82, 2018.