IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 29, 2019, accepted June 30, 2019, date of publication July 8, 2019, date of current version September 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927400

Generalized Information-Theoretic Criterion for Multi-Label Feature Selection

WANGDUK SEO, DAE-WON KIM, (Member, IEEE), AND JAESUNG LEE
School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Jaesung Lee (curseor@cau.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant 2019R1C1C1008404, and in part by the Chung-Ang University Graduate Research Scholarship, in 2019.

ABSTRACT Multi-label feature selection that identifies important features from the original feature set of multi-labeled datasets has been attracting considerable attention owing to its generality compared to conventional single-label feature selection. The unimportant features are filtered by scoring the dependency of features to labels. In conventional multi-label feature filter studies, the score function is obtained by approximating a dependency measure such as joint entropy because direct calculation is often impractical due to the presence of multiple labels with limited training patterns. Although the efficacy of approximation can differ depending on the characteristics of the multi-label dataset, conventional methods presume a certain approximation method, leading to a degenerated feature subset if the presumed approximation is inappropriate for the given dataset. In this study, we propose a strategy for selecting an approximation among a series of approximations depending on the number of involved variables and consequently instantiate a score function based on the chosen approximation. The experimental results demonstrate that the proposed strategy and score function outperform conventional multi-label feature selection methods.

INDEX TERMS Machine learning, multi-label learning, multi-label feature selection, information entropy.

I. INTRODUCTION

Multi-label feature selection (MLFS) identifies a subset of important features dependent on a given label set [1], [2]. Because the redundant features of input data can be reduced and possibly improve the accuracy of latter learning methods or classifiers [1], [3], [4], MLFS is regarded as a promising technique within the machine learning community [5]–[7]. Given input data with original feature set F and label set L where $|L| = t$, the goal of MLFS is to identify a feature subset $S \subset F$ with $n \ll |F|$ features that have the largest dependency on multiple labels [8]. Because S should concurrently support multiple labels using only n features [9], MLFS is a generalized method of a conventional single-label feature selection that can be applied to several varied problems for both single and multi-label datasets.

Recently, many studies have reported effective methods for MLFS that select most important features ranked by its corresponding criterion [10], [11]. Among the criteria considered from this field [12], [13], the dependency between a feature subset S and L can be measured using mutual information [1], [8], [14]–[17].

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

where $H(X) = -\sum P(X) \log P(X)$ is the joint entropy of the involved variable set X. The equation indicates that the estimation of high dimensional joint probability is required to obtain the values of $H(S, L)$ and $H(L)$.

The joint entropy of X can be approximated by obtaining the sum of entropies of subsets of X [18]. In most studies [1], [17], the subset size is fixed at a specific number k to approximate the joint entropy in a practical way, called a k-cardinality entropy approximation, wherein k is usually fixed at two. However, applying an optimal k-cardinality approximation to calculate an accurate estimation of the joint entropy between high-dimensional labels and features is practically unfeasible owing to insufficient patterns in a dataset [9]. Furthermore, the required computational cost increases considerably when k increases as that may exceed computational limits. For this reason, the optimal value of cardinality k is dependent to datasets and computational resources. As the optimal k is unknown until the actual result is verified, one of the solutions is to give a generalized approximation method that can instantiated various

The associate editor coordinating the review of this article and approving it for publication was Jad Nasreddine.

k-cardinality based criterion to user. One reason for this is that the user can obtain different performances based on varied k-cardinality and chose the performance that was better. In addition, the user can apply k-cardinality approximation as it is suitable for computational capacity. Therefore, we proposed a novel approximation generalized to cardinality k that can be selected to deal with various datasets. We generalized the approximation to calculate the entropy in terms of k and instantiate it to the k-cardinality entropy-based criterion specified by the users. Based on the derivation, a new criterion for MLFS problem is proposed. Finally, the main contributions of this paper are as follows:

- We proposed a novel MLFS based on the general form of entropy calculation.
- We compared the performance of our proposed method by varying the cardinality of approximation to search for a practical approximation.
- Comprehensive experimental results and statistical analysis proved the effectiveness of proposed method.

The remainder of this paper is organized as follows: Conventional MLFS methods are reviewed in Section II. In Section III, incremental search to navigate the optimal feature subset is introduced and a new generalized approximation technique is proposed. In Section IV, our analysis shows that the derived criterion is able to identify an effective feature subset for improving multi-label learning accuracy. Finally, our conclusions are given in Section V.

II. RELATED WORK

In this section, we briefly review related works on MLFS methods. Traditional feature selection for both single and multi-label learning involves two major methods; the filter method and the wrapper method. Both methods are aimed at selecting relevant features to search optimal feature subsets. The filter method evaluates the quality of features through mathematical criteria such as information theory [9], $l_{2,1}$-norm function optimization [19], quadratic optimization [17], and graph theory [20]. In contrast, the wrapper method evaluates feature subsets using classifiers and improves them. Generally, the wrapper method incurs a higher computational cost than the filter method, and selected feature subset by the wrapper method is dependent on the classifier used. Owing to its effectiveness and efficiency, the filter method is the most popular approach for MLFS problems [21]. Among filter-based MLFS, one of the major trends is transforming multi-label datasets to single-label datasets [2], [22]. This approach has advantages in terms of ease of use [23] because conventional single-label feature selection has been widely studied. For example, multiple labels can be transformed into a single label using the power-set of labels [24] wherein each combination of label subsets is assigned to a new class. However, this approach can cause overfitting or data imbalance issues owing to the creation of too many classes. In contrast, filter methods calculating the quality of features by relation between features and labels directly were proposed. Among various mathematical approaches, the effectiveness of information theory-based filter methods have been reported [1], [11], [13]. Pairwise multi-label utility (PMU) [9] was proposed based on mutual information that calculates the second-cardinality interaction of features and labels. A criterion based on Max-dependency and Min-redundancy (MDMR) [1] combines mutual information with max-dependency and min-redundancy. However, these information theory-based filter methods can suffer from inaccurate entropy estimation owing to specified second-cardinality entropy approximation that does not consider the characteristic of datasets. As it is not guaranteed that all datasets' optimal k-cardinality is second-cardinality, we need to investigate another k-cardinality entropy approximation-based criterion to obtain better performance.

III. PROPOSED METHOD

A. INCREMENTAL SEARCH

To identify the optimal feature subset that maximizes (1), feature selection method has to evaluate all possible $\binom{|F|}{n}$ candidate feature subsets because features in F can be dependent on each other [25]. As this exhaustive search strategy easily becomes an impractical task with a large number of $|F|$, an incremental search strategy can be considered to circumvent the prohibitive computational cost. In this manner, starting from $S = \{\emptyset\}$, the selected feature subset can be obtained by adding a feature $f^+ \in \{F - S\}$ sequentially that maximizes (2) until S is composed of desirable number of features n.

$$J = M(S, f^+; L) - M(S; L) \tag{2}$$

As the value of $M(S; L)$ in (2) is uninfluential in determining superiority among a new candidate feature f^+, (2) can be simplified as

$$\begin{aligned} J &\propto M(S, f^+; L) = H(S, f^+) - H(S, f^+, L) + H(L) \\ &\propto H(S, f^+) - H(S, f^+, L) \end{aligned} \tag{3}$$

B. APPROXIMATION OF JOINT ENTROPY

Let X' be the power set of X where $|X| = n$, and $X'_k = \{e | e \in X', |e| = k\}$ where $k \leq n$. Then the sum of k-cardinality entropy is defined as follows [9].

Definition 1: Sum of the k-cardinality entropy.

$$U_k(X) = \sum_{Y \in X_k} H(Y) \tag{4}$$

Based on Definition 1, Han's inequality can be rewritten as Proposition 1 [9], [26], [27].

Proposition 1: k-cardinality representation of Han's inequality.

$$H(X) \leq \frac{1}{n-1} U_{n-1}(X') \tag{5}$$

Proposition 1 indicates the upper bound of joint entropy involving n variables can be represented using a series of entropy terms involving $n - 1$ variables with a coefficient

related to the cardinality of X. Thus, the cardinality of involved variables is reduced compared to X. In addition, because each term in $U_{n-1}(X')$ is an entropy term, the upper bound of each term can also be obtained. By aggregating the upper bounds for each entropy term involving $n-1$ variables, the number of involved variables can be further reduced by applying Proposition 1 repeatedly to each entropy term in $U_{n-1}(X')$. To achieve this, an upper bound for $U_k(X')$ that deals the entropy terms involving variable subsets with k cardinality, can be written as Lemma 1.

Lemma 1: Upper bound of $U_k(X')$.

$$U_k(X') \leq \left(\frac{n-(k-1)}{k-1}\right) U_{k-1}(X') \tag{6}$$

Proof: Detailed derivation is provided in the work of [9]. Thus, Proposition 1 is a special case of Lemma 1 when $k = n$. □

By applying Lemma 1 repeatedly, a series of inequalities can be instantiated as follows.

$$H(X) \leq \frac{1}{n-1} U_{n-1}(X') \leq \frac{1}{n-1} \cdot \frac{2}{n-2} U_{n-2}(X') \leq \cdots$$

Thus, Lemma 2 that represents the series of upper bound for $H(X)$ can be obtained as follows.

Lemma 2: Series of upper bound for $H(X)$.

$$H(X) \leq \alpha_1 U_{n-1}(X') \leq \cdots \leq \alpha_{n-1} U_1(X') \tag{7}$$

where the coefficient α_m $1 \leq m \leq n-1$ is written as

$$\alpha_m = \prod_{i=1}^{m} \frac{i}{n-i} \tag{8}$$

Proof: For each $U_k(X')$ term, the value k is reduced once each time the upper bound is obtained, with entropy terms involving $k-1$ variables. Thus, it can be represented as $U_{k-1}(X')$ with a coefficient written as

$$U_k(X') \leq \beta_m U_{k-1}(X') \tag{9}$$

Because $m + k - 1 = n$, β_m can be rewritten as

$$\beta_m = \frac{n-(k-1)}{k-1} = \frac{n-(n-m)}{n-m} = \frac{m}{n-m} \tag{10}$$

The equation indicates that each process obtaining the upper bound implies a coefficient $m/(n-m)$. As the coefficient implied from each process is multiplied cumulatively, this can be formulated as follows.

$$\alpha_m = \prod_{i=1}^{m} \beta_i \tag{11}$$

By combining the sum of k-cardinality entropy and (11), Lemma 2 is obtained. □

Suppose that $|X| = 5$ for example, then the series of upper bound can be instantiated as

$$\begin{aligned} H(X) &\leq \frac{1}{4} U_4(X') \leq \cdots \leq \left(\frac{1}{4} \cdot \frac{2}{3} \cdot \frac{3}{2} \cdot \frac{4}{1}\right) U_1(X') \\ &= H(X) \leq \frac{1}{4} U_4(X') \leq \frac{1}{6} U_3(X') \leq \frac{1}{4} U_2(X') \leq U_1(X') \end{aligned}$$

Thus, the right most hand term of Lemma 2 indicates Shannon's intrinsic upper bound as

$$H(X) \leq \sum_{x \in X} H(x)$$

Lemma 2 indicates the value of each upper bound is monotonically increased as growing number of m because $H(X)$ and all the $U_k(X')$ terms are greater than or equal to zero. Moreover, Lemma 2 allow the approximation of $H(X)$ based on a series of entropy terms involving a desirable cardinality of variable subsets to ensure a reliable probability estimation according to the characteristics of given dataset. Suppose that the value m is chosen. Then the upper bound according to m can be written as Theorem 1.

Theorem 1: Upper bound of $H(X)$ with a specific cardinality.

$$H(X) \leq \left(\prod_{i=1}^{m} \frac{i}{n-i}\right) U_{n-m}(X') \tag{12}$$

where m denotes the number of times we obtained the upper bound of $H(X)$ sequentially.

Proof: From (8), the coefficient α_m for each term representing the sum of k-cardinality entropy was obtained. In addition, because the cardinality of the involved variable subsets decreased one by one from n as the number of times we obtained the upper bound increases, this can be formulated as $U_{n-m}(X')$. □

It will be convenient and also intuitive if Theorem 1 is rewritten based on the sum of k-cardinality entropy. As a result, Theorem 1 can be represented as Corollary 1.

Corollary 1: k-cardinality upper bound of $H(X)$.

$$H(X) \leq \left(\prod_{i=1}^{n-k} \frac{i}{n-i}\right) U_k(X') \tag{13}$$

Proof: Because the subscript $n-m$ is replaced by k, $m = n-k$. By rewriting m using n and k, Corollary 1 can be obtained. □

One last issue is related to the calculation of α_{n-k} because it incurs a series of multiplications to obtain the actual value when k is a small value. In this case, the calculation of α_{n-k} can be simplified, resulting in Corollary 2 as follows.

Corollary 2: Efficient calculation of coefficients.

$$H(X) \leq \left(\prod_{i=1}^{b} \frac{i}{n-i}\right) U_k(X') \tag{14}$$

where $b = \min(n-k, k-1)$.

Proof: Based on (8) and (13), the inverse of the coefficient is written as

$$\begin{aligned} \frac{1}{\alpha_{n-k}} &= \prod_{i=1}^{n-k} \frac{n-i}{i} = \frac{(n-1)\cdots(n-(n-k))}{(n-k)!} \\ &= \frac{(n-1)!}{(n-k)! \cdot (n-(n-k)-1)!} = \binom{n-1}{n-k} \end{aligned} \tag{15}$$

Because $\binom{n-1}{n-k} = \binom{n-1}{k-1}$ by the binomial theorem, (15) can be rewritten as

$$\frac{1}{\alpha_{n-k}} = \binom{n-1}{n-k} = \binom{n-1}{k-1} = \frac{1}{\alpha_{k-1}} \tag{16}$$

Because the value α_{n-k} is the same as α_{k-1}, the algorithm calculates α_{k-1} to obtain the value of α_{n-k} instead of calculating it directly if it reduces the calculation effort. □

Corollary 2 indicates symmetricity among coefficients. Suppose the algorithm chooses $k = 2$ to obtain the upper bound involving variable sets of cardinality two. Then the algorithm is able to obtain the value of α_{n-2} by calculating $\alpha_{k-1} = \alpha_1$ that incurs fewer number of multiplications from the calculation than that of α_{n-2}.

C. INSTANTIATION OF CRITERION

Based on Corollary 1, the upper bound involving variable subsets of k cardinality can be easily obtained. Suppose that a reliable joint entropy calculation can be achieved when $k = 2$ and the algorithm tries to calculate joint entropy terms in (3) by approximating them using their upper bounds. In this case, $H(X)$ is approximated as

$$H(X) \approx \left(\prod_{i=1}^{n-2} \frac{i}{n-i}\right) U_2(X') \tag{17}$$

Owing to Corollary 2, (17) can be simplified as

$$H(X) \approx \left(\prod_{i=1}^{1} \frac{i}{n-i}\right) U_2(X') = \frac{1}{n-1} U_2(X') \tag{18}$$

Based on (18), J that involves high-dimensional joint entropy calculation is approximated as

$$\begin{aligned} J &\propto H(S, f^+) - H(S, f^+, L) \\ &\approx \underbrace{\frac{1}{|S|} U_2(\{S, f^+\}')}_{\text{Part 1}} - \underbrace{\frac{1}{|S|+|L|} U_2(\{S, f^+, L\}')}_{\text{Part 2}} \end{aligned} \tag{19}$$

Part 1 in (19) can be rewritten as [18]

$$\frac{1}{|S|} U_2(\{S, f^+\}') = \underbrace{\frac{1}{|S|} U_2(S')}_{\text{Part 3}} + \frac{1}{|S|} U_2(f^+ \times S') \tag{20}$$

where $\times$ denotes the Cartesian products between two variable sets. Because Part 3 is a constant term for evaluating all candidate features in $\{F - S\}$, (20) can be simplified as

$$\frac{1}{|S|} U_2(\{S, f^+\}') \propto \frac{1}{|S|} U_2(f^+ \times S') = \frac{1}{|S|} \sum_{f \in S} H(f^+, f) \tag{21}$$

Subsequently, Part 2 in (19) can be rewritten as

$$\begin{aligned} &\frac{1}{|S|+|L|} U_2(\{S, f^+, L\}') \\ &= \frac{1}{|S|+|L|} (U_2(f^+ \times S') + U_2(f^+ \times L')) \\ &\quad + \underbrace{\frac{1}{|S|+|L|} (U_2(S') + U_2(L') + U_2(S' \times L'))}_{\text{Part 4}} \end{aligned} \tag{22}$$

Because Part 4 is a constant term, (22) can be simplified as

$$\begin{aligned} &\frac{1}{|S|+|L|} U_2(\{S, f^+, L\}') \\ &\propto \frac{1}{|S|+|L|} (U_2(f^+ \times S') + U_2(f^+ \times L')) \\ &= \frac{1}{|S|+|L|} \left(\sum_{f \in S} H(f^+, f) + \sum_{l \in L} H(f^+, l) \right) \end{aligned} \tag{23}$$

By combining (21) and (23), J is rewritten as

$$\begin{aligned} \tilde{J} \propto \underbrace{\frac{1}{|S|} \sum_{f \in S} H(f^+, f) - \frac{1}{|S|+|L|} \sum_{f \in S} H(f^+, f)}_{\text{Part 5}} \\ - \frac{1}{|S|+|L|} \sum_{l \in L} H(f^+, l) \end{aligned} \tag{24}$$

Part 5 in (24) can be simplified as

$$\begin{aligned} &\frac{1}{|S|} \sum_{f \in S} H(f^+, f) - \frac{1}{|S|+|L|} \sum_{f \in S} H(f^+, f) \\ &= \left(\frac{1}{|S|} - \frac{1}{|S|+|L|} \right) \sum_{f \in S} H(f^+, f) \\ &= \frac{1}{|S|+|L|} \cdot \frac{|L|}{|S|} \sum_{f \in S} H(f^+, f) \end{aligned} \tag{25}$$

By replacing Part 5 in (24) using (25), $\tilde{J}$ is simplified as

$$\tilde{J} \propto \underbrace{\frac{1}{|S|+|L|}}_{\text{Part 6}} \left(\frac{|L|}{|S|} \sum_{f \in S} H(f^+, f) - \sum_{l \in L} H(f^+, l) \right) \tag{26}$$

Because Part 6 in (26) is a constant term, $\tilde{J}$ is simplified as

$$\begin{aligned} \tilde{J} &\propto \frac{|L|}{|S|} \sum_{f \in S} H(f^+, f) - \sum_{l \in L} H(f^+, l) \\ &\propto |L| \sum_{f \in S} H(f^+, f) - |S| \sum_{l \in L} H(f^+, l) \end{aligned} \tag{27}$$

As a result, the criterion J is approximated as

$$\tilde{J} = \underbrace{|L| \sum_{f \in S} H(f^+, f)}_{\text{Part 7}} - \underbrace{|S| \sum_{l \in L} H(f^+, l)}_{\text{Part 8}} \tag{28}$$

The number of entropy terms in Part 7 is $|S|$ and that in Part 8 is $|L|$. The criterion $\tilde{J}$ does not tell how the first feature should be selected when $S = \{\emptyset\}$ because the approximator described in (18) assumes that there are at least two variables in X. For this case where $|S| = 0$, $\tilde{J}$ can be instantiated as follows.

$$\begin{aligned} J &\propto H(S, f^+) - H(S, f^+, L) \\ &= \underbrace{H(f^+)}_{\text{Part 9}} - \underbrace{H(f^+, L)}_{\text{Part 10}} \end{aligned} \tag{29}$$

Part 9 in (29) does not need to be approximated. On the contrary, Part 10 is approximated as

$$H(f^+, L) \approx \frac{1}{|L|} U_2(\{f^+, L\}') = \frac{1}{|L|} U_2(f^+ \times L') + \underbrace{\frac{1}{|L|} U_2(L')}_{\text{Part 11}} \quad (30)$$

Because Part 11 is a constant term, (30) can be simplified as

$$H(f^+, L) \propto \frac{1}{|L|} U_2(f^+ \times L') = \frac{1}{|L|} \sum_{l \in L} H(f^+, l) \quad (31)$$

By replacing Part 10 using (31), the criterion is written as

$$\begin{aligned} \tilde{J} &\propto H(f^+) - \frac{1}{|L|} \sum_{l \in L} H(f^+, l) \\ &\propto |L| \cdot H(f^+) - \sum_{l \in L} H(f^+, l) \end{aligned} \quad (32)$$

Thus, the criterion J when $S = \{\emptyset\}$ is approximated as

$$\tilde{J} = |L| \cdot H(f^+) - \sum_{l \in L} H(f^+, l) \quad (33)$$

For the cases where $k > 2$, the criterion can be easily instantiated in a similar manner. It is worth mentioning that an equivalent form of (33) is written as

$$\begin{aligned} \tilde{J} &= |L| \cdot H(f^+) - \sum_{l \in L} H(f^+, l) \\ &= |L| \cdot H(f^+) - \sum_{l \in L} H(f^+, l) + \sum_{l \in L} H(l) - \sum_{l \in L} H(l) \\ &= \sum_{l \in L} M(f^+; l) - \sum_{l \in L} H(l) \propto \sum_{l \in L} M(f^+; l) \end{aligned} \quad (34)$$

which is the same criterion when $S = \{\emptyset\}$ for most of mutual information-based MLFS methods [1], [3], [8], [9], [17], [18] despite the calculation of (33) being more efficient.

D. RELATION TO PREVIOUS STUDY

In conventional mutual information-based MLFS methods, the criteria were described using mutual information terms. Because the instantiated criterion is closely related to mutual information as demonstrated in Section III-C, the characteristics of $\tilde{J}$ may be intuitively identified after writing it using mutual information terms. Let us rewrite Part 7 in (28) using mutual information terms as follows.

$$\begin{aligned} \sum_{f \in S} H(f^+, f) &= -|S| \cdot H(f^+) + \sum_{f \in S} H(f^+, f) - \sum_{f \in S} H(f) \\ &\quad + |S| \cdot H(f^+) + \sum_{f \in S} H(f) \\ &= -\sum_{f \in S} M(f^+; f) + |S| \cdot H(f^+) + \sum_{f \in S} H(f) \end{aligned} \quad (35)$$

TABLE 1. Standard characteristics of employed datasets.

Datasets	Patterns	Features	Labels	Distinct label sets
Arts	7,484	1,157	26	599
Enron	1,702	1,001	53	753
Entertainment	12,730	1,600	21	337
Genbase	662	1,186	27	32
Health	9,205	1,530	32	335
Llog	1,460	1,004	75	304
Medical	978	1,449	45	94
Scene	2,407	294	6	15

Next, Part 8 can be rewritten as follows.

$$\begin{aligned} \sum_{l \in L} H(f^+, l) &= -|L| \cdot H(f^+) + \sum_{l \in L} H(f^+, l) - \sum_{l \in L} H(l) \\ &\quad + |L| \cdot H(f^+) + \sum_{l \in L} H(l) \\ &= -\sum_{l \in L} M(f; l) + |L| \cdot H(f^+) + \sum_{l \in L} H(l) \end{aligned} \quad (36)$$

By replacing Part 7 and Part 8 using (35) and (36) respectively, $\tilde{J}$ can be rewritten as follows.

$$\begin{aligned} \tilde{J} &= |L| \left(-\sum_{f \in S} M(f^+; f) + |S| \cdot H(f^+) + \sum_{f \in S} H(f) \right) \\ &\quad - |S| \left(-\sum_{l \in L} M(f^+; l) + |L| \cdot H(f^+) + \sum_{l \in L} H(l) \right) \end{aligned} \quad (37)$$

By cancelling out $H(f^+)$ terms, (37) can be rewritten as

$$\begin{aligned} \tilde{J} = |S| \sum_{l \in L} M(f^+; l) - |L| \sum_{f \in S} M(f^+; f) \\ + \underbrace{|L| \sum_{f \in S} H(f) - |S| \sum_{l \in l} H(l)}_{\text{Part 12}} \end{aligned} \quad (38)$$

Because Part 12 is a constant term, (38) can be simplified as

$$\begin{aligned} \tilde{J} &\propto |S| \sum_{l \in L} M(f^+; l) - |L| \sum_{f \in S} M(f^+; f) \\ &\propto \sum_{l \in L} M(f^+; l) - \frac{|L|}{|S|} \sum_{f \in S} M(f^+; f) \end{aligned} \quad (39)$$

Suppose that $L = \{l\}$ in which leads to a single-label feature selection problem. Then, the criterion $\tilde{J}$ is simplified as

$$\tilde{J} = M(f^+; l) - \frac{1}{|S|} \sum_{f \in S} M(f^+; f)$$

which is one of well-known criterion for single-label feature selection problems owing to its effectiveness in addition to its intuitiveness [25]. Thus, (39) identifies f^+ that maximizes

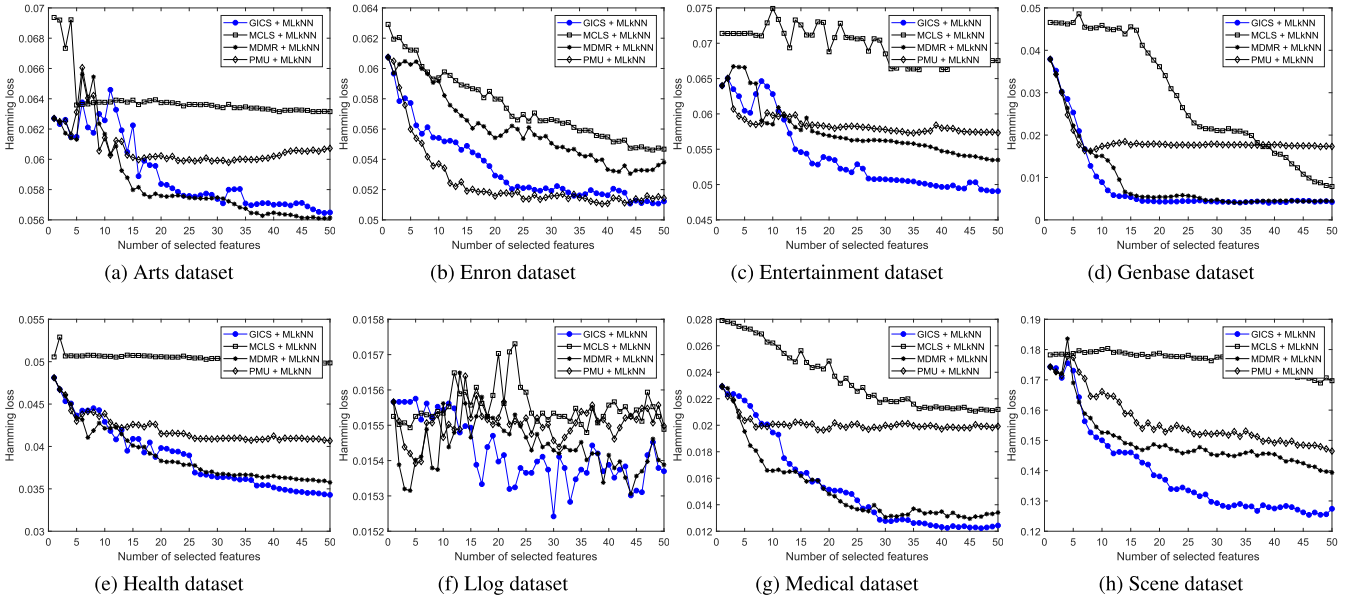
TABLE 2. Performance comparison (MLkNN) when $|S| = 30$.

Datasets	Multi-label accuracy (↑)				Hamming loss (↓)			
	GICS	MCLS	MDMR	PMU	GICS	MCLS	MDMR	PMU
Arts	**0.194±0.015**	0.022±0.011	0.164±0.006	0.131±0.022	**0.057±0.001**	0.063±0.001	0.057±0.001	0.060±0.001
Enron	**0.378±0.021**	0.300±0.033	0.312±0.034	0.376±0.008	0.052±0.002	0.057±0.002	0.055±0.002	**0.052±0.002**
Entertainment	**0.342±0.011**	0.062±0.073	0.228±0.012	0.187±0.021	**0.051±0.001**	0.068±0.007	0.056±0.001	0.058±0.001
Genbase	**0.952±0.017**	0.674±0.097	0.947±0.020	0.707±0.055	**0.004±0.001**	0.021±0.004	0.005±0.002	0.018±0.002
Health	**0.532±0.056**	0.243±0.167	0.528±0.012	0.398±0.055	**0.036±0.001**	0.050±0.001	0.037±0.001	0.041±0.001
Llog	**0.191±0.030**	0.162±0.016	0.172±0.013	0.165±0.021	**0.015±0.001**	0.016±0.000	0.015±0.000	0.015±0.000
Medical	**0.659±0.035**	0.347±0.070	0.638±0.043	0.371±0.024	**0.013±0.001**	0.022±0.002	0.013±0.002	0.020±0.001
Scene	**0.432±0.020**	0.043±0.015	0.320±0.026	0.310±0.028	**0.129±0.006**	0.177±0.003	0.145±0.005	0.152±0.003
Avg.Rank	**1.000**	4.000	2.125	2.875	**1.125**	4.000	2.125	2.750
Datasets	Ranking loss (↓)				Normalized coverage (↓)			
	GICS	MCLS	MDMR	PMU	GICS	MCLS	MDMR	PMU
Arts	**0.147±0.004**	0.177±0.006	0.152±0.003	0.160±0.004	**0.246±0.005**	0.274±0.007	0.250±0.004	0.260±0.005
Enron	0.095±0.004	0.106±0.005	0.098±0.005	**0.095±0.004**	0.274±0.008	0.290±0.011	0.278±0.011	**0.273±0.011**
Entertainment	**0.114±0.004**	0.161±0.013	0.129±0.004	0.133±0.002	**0.198±0.004**	0.242±0.014	0.212±0.005	0.215±0.003
Genbase	**0.007±0.003**	0.024±0.006	0.010±0.003	0.022±0.006	**0.059±0.006**	0.077±0.008	0.065±0.006	0.078±0.010
Health	**0.052±0.002**	0.083±0.003	0.055±0.002	0.063±0.002	**0.122±0.003**	0.156±0.004	0.126±0.004	0.136±0.003
Llog	0.148±0.012	0.153±0.010	**0.146±0.012**	0.153±0.013	0.194±0.015	0.199±0.013	**0.191±0.015**	0.198±0.015
Medical	**0.039±0.007**	0.066±0.011	0.048±0.011	0.082±0.013	**0.079±0.008**	0.109±0.013	0.090±0.013	0.126±0.015
Scene	**0.162±0.018**	0.356±0.016	0.183±0.012	0.207±0.017	**0.315±0.014**	0.478±0.014	0.333±0.010	0.353±0.015
Avg.Rank	**1.250**	3.875	2.000	2.875	**1.250**	3.750	2.000	3.000

TABLE 3. Performance comparison (MLNB) when $|S| = 30$.

Datasets	Multi-label accuracy (↑)				Hamming loss (↓)			
	GICS	MCLS	MDMR	PMU	GICS	MCLS	MDMR	PMU
Arts	**0.213±0.013**	0.038±0.015	0.201±0.006	0.130±0.017	**0.057±0.001**	0.067±0.002	0.063±0.001	0.062±0.001
Enron	**0.371±0.015**	0.293±0.034	0.267±0.017	0.345±0.016	**0.058±0.003**	0.070±0.006	0.117±0.006	0.068±0.005
Entertainment	**0.338±0.012**	0.048±0.045	0.255±0.033	0.192±0.019	**0.053±0.001**	0.068±0.001	0.069±0.006	0.073±0.001
Genbase	**0.957±0.018**	0.709±0.053	0.911±0.018	0.694±0.041	**0.005±0.002**	0.023±0.004	0.009±0.002	0.018±0.002
Health	**0.565±0.011**	0.325±0.084	0.529±0.009	0.388±0.011	**0.036±0.001**	0.051±0.001	0.042±0.002	0.044±0.002
Llog	**0.221±0.021**	0.187±0.025	0.159±0.077	0.132±0.071	**0.02±0.002**	0.031±0.005	0.046±0.005	0.039±0.007
Medical	**0.712±0.030**	0.409±0.039	0.539±0.026	0.399±0.026	**0.012±0.001**	0.030±0.005	0.028±0.002	0.024±0.002
Scene	**0.531±0.013**	0.289±0.011	0.43±0.013	0.470±0.024	**0.158±0.006**	0.286±0.008	0.231±0.008	0.163±0.010
Avg.Rank	**1.000**	3.375	2.500	3.125	**1.000**	3.375	3.000	2.625
Datasets	Ranking loss (↓)				Normalized coverage (↓)			
	GICS	MCLS	MDMR	PMU	GICS	MCLS	MDMR	PMU
Arts	**0.145±0.018**	0.182±0.019	0.159±0.016	0.159±0.018	**0.243±0.016**	0.279±0.018	0.259±0.014	0.257±0.017
Enron	**0.092±0.007**	0.111±0.008	0.147±0.009	0.115±0.010	**0.267±0.011**	0.299±0.013	0.365±0.016	0.307±0.012
Entertainment	**0.110±0.002**	0.156±0.008	0.138±0.004	0.147±0.004	**0.195±0.003**	0.239±0.009	0.226±0.005	0.234±0.004
Genbase	**0.029±0.026**	0.046±0.025	0.031±0.025	0.044±0.026	**0.077±0.025**	0.095±0.025	0.080±0.024	0.096±0.025
Health	**0.080±0.028**	0.111±0.028	0.086±0.028	0.093±0.027	**0.149±0.026**	0.182±0.026	0.159±0.025	0.164±0.024
Llog	**0.166±0.020**	0.176±0.025	0.167±0.022	0.187±0.021	0.209±0.020	0.219±0.025	**0.207±0.023**	0.231±0.022
Medical	0.078±0.023	0.090±0.023	**0.073±0.025**	0.108±0.024	0.113±0.023	0.126±0.023	**0.106±0.025**	0.146±0.026
Scene	**0.120±0.010**	0.330±0.011	0.155±0.009	0.171±0.022	**0.280±0.008**	0.455±0.009	0.309±0.008	0.323±0.019
Avg.Rank	**1.125**	3.500	2.125	3.250	**1.250**	3.375	2.125	3.250

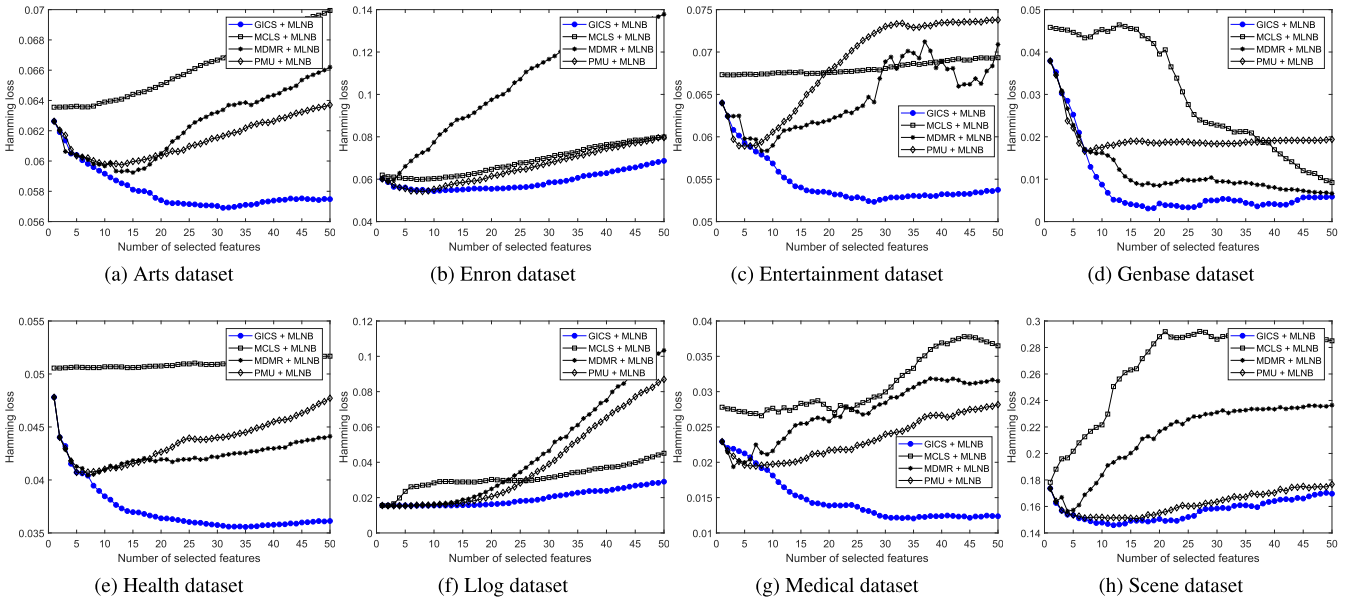
the dependency between f^+ and labels in L while minimizing the dependency between f^+ and features in S because it will imply a large and a small value of Part 7 and Part 8, respectively. We named our method as Generalized Information-theoretic Criterion for MLFS (GICS), in which the procedural step for identifying the final feature subset can be described as Algorithm 1. The time complexity of proposed GICS can be written as $O(|F| \cdot |L| + |F| \cdot |S|)$. It would be worth mentioning that the calculation results of joint entropy terms between features and labels in Line 4 can be reused from the



(a) Arts dataset (b) Enron dataset (c) Entertainment dataset (d) Genbase dataset

(e) Health dataset (f) Llog dataset (g) Medical dataset (h) Scene dataset

FIGURE 1. Hamming loss performance of Multi-label Nearest Neighbor classifer on eight datasets according to feature subsets using GICS and conventional feature selection methods: MCLS, MDMR, and PMU.



(a) Arts dataset (b) Enron dataset (c) Entertainment dataset (d) Genbase dataset

(e) Health dataset (f) Llog dataset (g) Medical dataset (h) Scene dataset

FIGURE 2. Hamming loss performance of Multi-label Naive Bayes classifier on eight datasets according to feature subsets using GICS and conventional feature selection methods: MCLS, MDMR, and PMU.

calculation of Line 6, leading to the acceleration of procedure with additional memory usage.

IV. EXPERIMENTAL RESULTS

A. DATASETS AND EVALUATION

We experimented on eight multi-label datasets from various domains. Enron is a subset of the Enron email corpus wherein each email document can be categorized according to multiple objectives [28]. Genbase comes from the biological domain that includes the functions of genes and proteins [29]. Medical dataset is sampled from a large corpus of suicide letters obtained from natural language preprocessed clinical free text [30]. Language Log (LLog) was generated from text mining, wherein each feature corresponds to the occurrence of a word and each label represents the relevancy of each text pattern to a specific subject [31]. The Scene dataset is related to the semantic indexing of still scenes, where each scene may contain multiple objects [32]. The remaining three datasets (Arts, Entertainment, Health) comes from the Yahoo dataset collection. Table 1 shows the standard statistics of the datasets

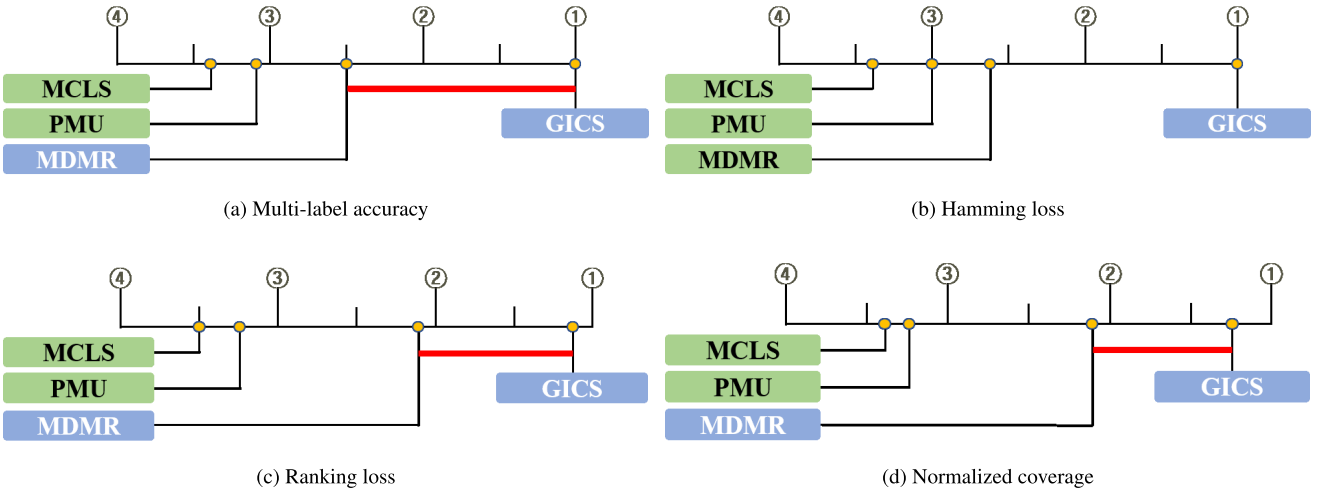


FIGURE 3. Hamming loss performance of MLNB according to feature subsets using GICS with varying k.

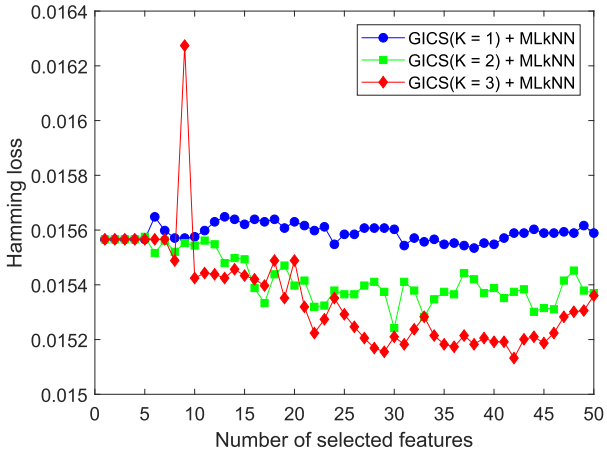


FIGURE 4. Hamming loss performance of MLkNN on the Llog dataset according to feature subsets using GICS with varying k.

employed in our experiments such as the number of patterns, features, labels, and distinct label sets [31], [33].

We compared our proposed GICS to three conventional MLFS methods: Manifold-based Constraint Laplacian Score (MCLS) [10], MDMR [1], and PMU [9]. Different from the embedded methods such as Lasso Regularization and Decision Tree, the proposed method is independent of the classification process, indicating that the subsequent multi-label classifier can be varied. To evaluate the effectiveness of the selected feature subsets, two well-known multi-label classifiers were employed; Multi-label Naive Bayes (MLNB) [34] and Multi-label Nearest Neighbor (MLkNN) with $k = 10$ [35]. Experiments were repeated ten times on each dataset conducted by a holdout cross-validation; 80% of patterns were randomly chosen as a training set and the rest 20% were used as the test set to obtain the classification performance. The predicted label from the test set by classifiers are measured using four evaluation metrics: multi-label accuracy, Hamming loss, ranking loss, and normalized coverage. Higher values of multi-label accuracy and lower values of the remainder (Hamming loss, ranking loss, normalized coverage) indicate good classification performance. A detailed definition of each evaluation measure is described in [4], [33].

B. COMPARISON RESULTS

Figs. 1 and 2 show the Hamming loss performance of multi-label classifiers depending on each feature subset selected by the compared methods. In Fig. 1a, the horizontal axis represents the number of selected features, and the vertical axis represents the corresponding Hamming loss value. In Fig. 1, MLkNN is used to predict labels. GICS outperformed the compared methods in the Entertainment, Llog, and Scene datasets when the size of the feature subset is larger than 15. In Fig. 1d, GICS and MDMR's Hamming loss performances are converged when the size of feature subset is larger than 15 in the Genbase dataset. However GICS shows faster convergence than MDMR. In Fig. 2, MLNB is used to predict labels. GICS outperformed the compared methods in all datasets.

Tables 2 and 3 show the performance values depending on four evaluation measures for each method when the number of selected features is 30 and the average rank of each methods on all datasets. For each evaluation measure, a down arrow symbol in the Hamming loss, ranking loss, and normalized coverage indicates a smaller value for the corresponding performance value, demonstrating better quality of classification results. The best value along with each dataset is marked in bold text. The results indicate that GICS outperformed the other compared methods for most datasets. Even if GICS did not show the best performance such as multi-label accuracy in Enron datasets, GICS shows second-best results in all datasets that did not show the best performance.

Algorithm 1 Proposed GICS

1: **Input:** F, L, n; ▷ Number of features to be selected, n
2: **Output**: S; ▷ Selected feature subset, S
3: $S \leftarrow \{\phi\}$;
4: Include $f^+ \in F$ to S that maximizes Eq. (33);
5: **while** $|S| < n$ **do**
6: Include $f^+ \in \{F - S\}$ to S that maximizes Eq. (28);
7: **end while**

TABLE 4. Friedman statistics $F_F(c = 4, N = 8)$ and critical value in terms of each evaluation measure by the MLNB.

Evaluation measure	F_F	Critical values ($\alpha = 0.05$)
Multi-label accuracy	14.9607	3.0724
Hamming loss	13.3636	
Ranking loss	17.8888	
Normalized coverage	10.7777	

We conducted statistical tests to confirm GICS's performance [36]. Table. 4 shows the results of the Friedman test in terms of all performance measures at a significance level of $\alpha = 0.05$ and the critical value for each evaluation measure. The Friedman statistics F_F are larger than the critical values in terms of all evaluation measures. Thus, the null hypothesis, which means the proposed methods and other conventional methods show no significant difference, is rejected. After the Friedman test, we performed the Bonferroni-Dunn post-hoc test. Fig. 3 shows the Critical Distance diagrams for each evaluation measure indicating the average ranks of each methods. In the diagrams, the conventional methods are connected with the GICS with a thick line, indicating no significant difference in performance. In terms of multi-label accuracy, ranking loss, and normalized coverage, the GICS shows statistical superiority in the performances to MCLS, PMU. In terms of Hamming loss, the GICS shows statistical superiority compared to all conducted conventional methods.

C. CARDINALITY ANALYSIS

Through GICS, various criteria can be created based on what cardinality k is chosen depending on the characteristics of the datasets. Although our primary criterion is derived based on $k = 2$, we conducted the final experiment to validate the classification performance with varying k and obtain an optimal option of k because each criterion varied by k can output different feature subsets. Figs. 4 shows the Hamming loss performance by the MLkNN of feature subsets each selected by the criterion based on $k = 1, 2$, and 3 on the Llog dataset. Considering the criterion of $k = 1$ is

$$\tilde{J} = H(f^+)$$

which selects features as a manner of unsupervised feature selection. The experimental results indicate that the Hamming loss performance of the criterion based on $k = 2$ and 3 outperformed the criterion based on $k = 1$ after more than ten features are selected. The Hamming loss performance of the criterion based on $k = 1$ is stagnates around 0.0156 when the number of selected features increases. Both criterion based on $k = 2$ and 3 do not show a significant difference in the Hamming loss performance when the number of the selected features is lower than 20. After more than 20 features are selected, the performance of the criterion based on $k = 3$ is consistently better than that based on $k = 2$. However, with regard to the execution time, the criterion based on $k = 3$ consumes considerably more time and memory; for example, in the Llog dataset, the criterion based on $k = 3$ consumes approximately 95.66 min for each iteration, whereas that based on $k = 2$ consumes 1.63 min from MATLAB on the Intel Core i7-7700 (3.6GHz) with 32GB RAM.

V. CONCLUSION

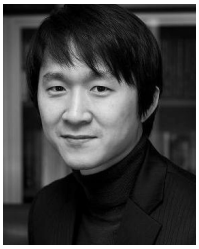
In this study, we proposed a cardinality generalized entropy approximation method wherein cardinality can be chosen by users. There is a trade-off that using various cardinality based criteria can help us find better feature subsets but that it requires higher computational cost. Our theoretical analysis shows that the derived criterion is closely related to the well-known criterion for conventional single-label feature selection problem. The experimental results demonstrate the effectiveness of our proposed method and proved that our proposed method can obtain a good feature subset.

REFERENCES

[1] 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, Nov. 2015.
[2] 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, Mar. 2016.
[3] J. Lee and D.-W. Kim, "Memetic feature selection algorithm for multi-label classification," *Inf. Sci.*, vol. 293, pp. 80–96, Feb. 2015.
[4] M.-L. Zhang and L. Wu, "LIFT: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.
[5] S. M. Liu and J.-H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1083–1093, Feb. 2015.
[6] Y. Rao, "Contextual sentiment topic model for adaptive social emotion classification," *IEEE Intell. Syst.*, vol. 31, no. 1, pp. 41–47, Jan./Feb. 2016.
[7] Z. Sun, J. Zhang, L. Dai, C. Li, C. Zhou, J. Xin, and S. Li, "Mutual information based multi-label feature selection via constrained convex optimization," *Neurocomputing*, vol. 329, pp. 447–456, Feb. 2019.
[8] J. Lee and D.-W. Kim, "Fast multi-label feature selection based on information-theoretic feature ranking," *Pattern Recognit.*, vol. 48, no. 9, pp. 2761–2771, 2015.
[9] J. Lee and D.-W. Kim, "Mutual information-based multi-label feature selection using interaction information," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 2013–2025, 2015.
[10] R. Huang, W. Jiang, and G. Sun, "Manifold-based constraint Laplacian score for multi-label feature selection," *Pattern Recognit. Lett.*, vol. 112, pp. 346–352, Sep. 2018.
[11] Y. Lin, Q. Hu, J. Liu, J. Chen, and J. Duan, "Multi-label feature selection based on neighborhood mutual information," *Appl. Soft Comput.*, vol. 38, pp. 244–256, Jan. 2016.
[12] J. Xu and Q. Ma, "Multi-label regularized quadratic programming feature selection algorithm with Frank–Wolfe method," *Expert Syst. Appl.*, vol. 95, pp. 14–31, Apr. 2018.
[13] J. Lee and D.-W. Kim, "SCLS: Multi-label feature selection based on scalable criterion for large label set," *Pattern Recognit.*, vol. 66, pp. 342–352, Jun. 2017.
[14] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[15] G. Doquire and M. Verleysen, "Mutual information-based feature selection for multilabel classification," *Neurocomputing*, vol. 122, pp. 148–155, Dec. 2013.
[16] J. Lee, H. Lim, and D. W. Kim, "Approximating mutual information for multi-label feature selection," *Electron. Lett.*, vol. 48, no. 15, pp. 929–930, Jul. 2012.
[17] H. Lim, J. Lee, and D.-W. Kim, "Multi-label learning using mathematical programming," *IEICE Trans. Inf. Syst.*, vol. 98, no. 1, pp. 197–200, 2015.
[18] J. Lee and D.-W. Kim, "Feature selection for multi-label classification using multivariate mutual information," *Pattern Recognit. Lett.*, vol. 34, no. 3, pp. 349–357, 2013.
[19] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, vol. 23, 2010, pp. 1813–1821.
[20] X. Kong and P. S. Yu, "gMLC: A multi-label feature selection framework for graph classification," *Knowl. Inf. Syst.*, vol. 31, no. 2, pp. 281–305, 2012.
[21] S. Kashef, H. Nezamabadi-Pour, and B. Nikpour, "Multilabel feature selection: A comprehensive review and guiding experiments," *WIREs Data Min. Knowl. Discovery*, vol. 8, no. 2, p. e1240, 2018.
[22] J. Read, "A pruned problem transformation method for multi-label classification," in *Proc. New Zealand Comput. Sci. Res. Student Conf.*, Christchurch, New Zealand, Apr. 2008, pp. 143–150.
[23] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.
[24] 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," *Electron. Notes Theor. Comput. Sci.*, vol. 292, pp. 135–151, Mar. 2013.
[25] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[26] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Inf. Control*, vol. 36, no. 2, pp. 133–156, 1978.
[27] M. Madiman and P. Tetali, "Information inequalities for joint distributions, with interpretations and applications," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2699–2713, Jun. 2010.
[28] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *Proc. Eur. Conf. Mach. Learn.*, Pisa, Italy, Sep. 2004, pp. 217–226.
[29] S. Diplaris, G. Tsoumakas, P. A. Mitkas, and I. Vlahavas, "Protein classification with multiple algorithms," in *Advances in Informatics*, vol. 3746. Nov. 2005, pp. 448–456.
[30] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, and W. Duch, "A shared task involving multi-label classification of clinical free text," in *Proc. Workshop BioNLP, Biol., Transl., Clin. Lang. Process.*, Prague, Czech, 2007, pp. 97–104.
[31] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, "Mulan: A java library for multi-label learning," *J. Mach. Learn. Res.*, vol. 12, pp. 2411–2414, Jun. 2011.
[32] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.
[33] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[34] M.-L. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive Bayes classification," *Inf. Sci.*, vol. 179, no. 19, pp. 3218–3229, 2009.
[35] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.
[36] J. Demšar, "Statistical comparisons of classifiersover multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, Jan. 2006.

WANGDUK SEO received the B.S. and M.S. degrees in computer science from Chung-Ang University, Seoul, South Korea, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interests include meta-heuristic optimization, multi-label learning, and feature selection.



DAE-WON KIM received the B.S. degree from Kyungpook National University, Daegu, South Korea, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology. He completed his Postdoctoral Research at the Korea Advanced Institute of Science and Technology. He is currently a Professor with the School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea. His research interests include advanced data mining algorithms with innovative applications to bioinformatics, music emotion recognition, educational data mining, affective computing, and robot interaction.

JAESUNG LEE received the B.S., M.S., and Ph.D. degrees in computer science from Chung-Ang University, Seoul, South Korea, in 2007, 2009, and 2013, respectively, where he is currently a Research Professor. His research interests include data mining with applications to affective computing and ambient intelligence. In theoretical domain, he also studies classification, feature selection, and especially multi-label learning with information theory.