Pattern Recognition Letters 34 (2013) 349–357

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Feature selection for multi-label classification using multivariate mutual information

Jaesung Lee, Dae-Won Kim *

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

ARTICLE INFO

Article history:
Received 3 April 2012
Available online 2 November 2012

Communicated by S. Sarkar

Keywords:
Multi-label feature selection
Multivariate feature selection
Multivariate mutual information
Label dependency

ABSTRACT

Recently, classification tasks that naturally emerge in multi-label domains, such as text categorization, automatic scene annotation, and gene function prediction, have attracted great interest. As in traditional single-label classification, feature selection plays an important role in multi-label classification. However, recent feature selection methods require preprocessing steps that transform the label set into a single label, resulting in subsequent additional problems. In this paper, we propose a feature selection method for multi-label classification that naturally derives from mutual information between selected features and the label set. The proposed method was applied to several multi-label classification problems and compared with conventional methods. The experimental results demonstrate that the proposed method improves the classification performance to a great extent and has proved to be a useful method in selecting features for multi-label classification problems.

1. Introduction

Multi-label classification is a challenging problem that emerges in several modern applications such as text categorization, gene function classification, and semantic annotation of images (Schapire and Singer, 2000; Sebastiani, 2002; Lewis et al., 2004; Diplaris et al., 2005; Boutell et al., 2004). As in the traditional classification problem, the performance of multi-label classification is strongly influenced by the quality of input features. Theoretically, a pattern may lose its distinguishment owing to the irrelevant or redundant features since the similarity of each pair of patterns in same class can be decreased (Watanabe, 1969). These features could cause additional problems of confusing the learning algorithm and leading to poor classification performance (Guyon and Elisseeff, 2003; Saeys et al., 2007).

Consequently, most recent research concerned with multi-label classification naturally employed feature selection techniques (Yang and Pedersen, 1997; Chen et al., 2007; Doquire and Verleysen, 2011; Trohidis et al., 2008). The feature selection is a task of selecting relevant features directly to preserve the internal meaning of given features as it is. This is an important constraint in some applications; for example, the task of gene function classification considers the classification accuracy as well as the biological analysis of the selected features (Diplaris et al., 2005). In the present study, we focused on the feature selection approach to improve the performance of multi-label classification while preserving the inherent meaning of given features.

To select a set of relevant features from given data set, some multi-label feature selection algorithms optimize a set of parameters during feature selection process to tune the kernel function of multi-label classifier (Gu et al., 2011). However, it frequently encounters exhaustive calculations to find an appropriate hyperspace using pairwise comparisons of patterns. This process should be done in each iterative feature selection step, so it is impractical in the viewpoint of computational cost. There is another way of treating multi-label learning; this approach converts the multi-label problems into traditional single-label multi-class problem, and then each feature is evaluated in terms of dependency to transformed new single-label (Chen et al., 2007; Trohidis et al., 2008). This is the most simple approach and provides a connection between single-label learning researches and novel multi-label learning. However, it causes subsequent problems, since multiple labels are transformed to a single label, so that newly created label inherently contains too many classes, leading to difficulty of learning (Read, 2008).

In this paper, we propose a mutual information based multi-label feature selection criterion. The characteristic of our proposed method is that it does not involve any type of transformation method – it selects an effective feature subset by maximizing the dependency between selected features and labels. To the best of our knowledge, it is the first time of proposing a feature filter criterion that takes into account label interactions in evaluating the dependency of given features without resorting to problem transformation. This paper is organized as follows: Section 2 gives a

* Corresponding author. Tel.: +82 2 820 5304.
E-mail address: dwkim@cau.ac.kr (D.-W. Kim).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.10.005

detailed description of conventional feature selection methods that require a transformation method. In Section 3, we propose our multi-label feature filter criterion. To achieve this, we decompose the calculation of high-dimensional entropy into a cumulative sum of multivariate mutual information. The performance of the proposed method is investigated with several evaluation measures for various multi-label data sets in Section 4. The discussion and conclusions are presented in Section 5.

2. Related work

Before reviewing conventional feature selection methods, we introduce some basic notations for the multi-label learning. Let $W \subset \mathbb{R}^d$ denote an input space that is constructed from d features, and patterns drawn from W are assigned to a certain label subset $\lambda \subseteq L$, where $L = \{l_1, \ldots, l_t\}$ is a finite set of labels with $|L| = t$. Thus, multi-label classification is the task of assigning unseen patterns to multiple labels. To solve the multi-label classification problem, an algorithm should take into account many labels concurrently. Popular multi-label learning algorithms first transform label sets into a single label (a process called problem transformation) and then solve the resultant problem (Tsoumakas and Katakis, 2007; Yang and Pedersen, 1997; Doquire and Verleysen, 2011). Similarly, feature selection steps in the problem transformation approach are as follows: (1) Transform the original multi-label data set into a single-label data set. (2) Assess each feature independently using a score evaluation method such as mutual information (MI) or χ^2 statistics. (3) Select the predefined top n features as input features for the multi-label classifier. We represent this process as problem transformation +score measure, a notation that will be used subsequently.

Chen et al. (2007) proposed an Entropy-based Label Assignment (ELA) that assigns weights to a multi-label pattern for different labels based on the label entropy. The ELA copies each pattern in accordance with the number of its labels, and then the inverse of the number of its labels is assigned as the weight of each pattern. So, each original pattern-labels pair (P_i, λ) is transformed to a set of patterns $T_i = \{(P_{i1}, l_1), \ldots, (P_{i|\lambda|}, l_{|\lambda|})\}$ with its weights $\frac{1}{|\lambda|}$ where $1 \leqslant i \leqslant |W|$ and $l_j \in \lambda$. Since patterns with too many labels blurred out from the training phase owing to the assignment of low weight to this pattern, they argued that the learning algorithm can avoid the overfitting problem originating from these patterns. Text categorization data sets were transformed by ELA, and three feature selection methods were then exhaustively applied to each transformed data set; two feature selection methods employed information gain and χ^2 statistics as their score measure, and the other one used an optimal orthogonal centroid feature selection method. Their empirical experiments indicate that any problem transformation method yielding a loss of information about dependency among labels may lead to poor classification performance, even though the classification performance was improved by feature selection methods.

The Label Powerset (LP) is applied to music information retrieval, specifically for recognizing six emotions that are simultaneously evoked by a music clip (Trohidis et al., 2008). It transforms a multi-label to a single-label by assigning each pattern's label set to a single class, so each pattern-labels pair (P_i, λ) is transformed to (P_i, c_i) where $c_i \in \{0, 1\}^t$, and 0 for $l_j \notin \lambda$ while 1 for $l_j \in \lambda$ where $1 \leqslant j \leqslant t$. Suppose a pattern P_i is assigned to l_1, l_2, and l_5 simultaneously, then the transformed pattern-class pair is represented as $(P_i, \{1, 1, 0, 0, 1\})$ where $t = 5$. The total number of classes is the total number of distinct label sets. χ^2 statistics is used to select effective features with the LP to improve the recognition performance of the multi-labeled music emotions. The results indicate that a feature selection method that evaluates the dependency of each feature by considering each label independently may not lead to better classification performance. They argued that the best classification performance can be achieved by using LP + χ^2, since the LP considers label correlations directly. Although the LP is able to provide an intuitive way of transforming and takes into account relationships between labels, it suffers from class size issues (Tsoumakas et al., 2011). If there are rarely observed labels in the label set, then the LP creates too many classes, causing overfitting and imbalance problems (Sun et al., 2009).

Read (2008) proposed the Pruned Problem Transformation (PPT) to improve the LP; patterns with too rarely occurring labels are simply removed from the training set by considering label sets with a predefined minimum occurrence τ. Doquire and Verleysen (2011) proposed a multi-label feature selection method using PPT to improve the classification performance of image annotation and gene function classification. First, the multi-label data set is transformed using the PPT method; next, a sequential forward selection is undertaken with the MI as the search criterion. Empirical results show that this gives better classification performance than PPT + χ^2 when multi-label k nearest neighbors is applied (Zhang and Zhou, 2007). They indicate that mutual information can be used as a good score measure for evaluating the dependencies among features and labels, leading to good classification performance. However, since patterns could be discarded from original data set, this is an irreversible transformation, in which there may be loss of class information. As a result, the performance of learning algorithms may be limited, since the parameter τ is generally unknown in practical situations.

The limitation in recent multi-label feature selection methods is that they require a problem transformation method for evaluating the dependency of given features. Since problem transformation converts the multi-label problem into single-label problem, this process could cause subsequent problems. For example, if transformed single-label is composed of too many classes, the performance of learning algorithm could be degraded. Moreover, if information loss occurs in the transformation process, the feature selection cannot take into account label relations. As a result, it is important to develop a feature selection method that considers multi-labels directly. Therefore, we investigate a mutual-information-based feature selection method that does not require any problem transformation. In the next section, we propose our multi-label feature selection method.

3. Multivariate mutual information for multi-label feature selection

The feature selection problem is to select a subset S composed of selected n features, from a set of features F ($n < d$), which jointly have the largest dependency on L. To solve the feature selection problem, we should find relevant features that contain as much discriminating power about the output labels L as possible. In this section, we derive our multi-label multivariate filter criterion from the equation of mutual information between feature set S and label set L. The mutual information between selected feature subset S and label set L can be represented as follows:

$$\begin{aligned} I(S;L) &= H(S) - H(S,L) + H(L) \\ &= H(\{f_1, \ldots, f_n\}) - H(\{f_1, \ldots, f_n, l_1, \ldots, l_t\}) + H(\{l_1, \ldots, l_t\}) \end{aligned} \tag{1}$$

Each $H(\cdot)$ term of Eq. (1) represents the joint entropy of an arbitrary number of variables, defined as

$$H(X) = -\sum P(X) \log P(X) \tag{2}$$

where $P(X)$ is a probabilistic mass function of given a set of variables X. The entropy is a measure for self-content of a variable

set, on the contrary, the mutual information focuses on shared information between variables. Note that if given variables are composed of continuous variables, we can use either the differential entropy to obtain the entropy of continuous variables, or the preprocessing (discretization) method to transform continuous variables into discrete (categorical) counterparts. For the sake of simplicity, we present the key notion of the proposed method by using categorical variables, in which each variable have finite number of categories or discretized values.

The entropy term requires a high-dimensional probability estimation of the given variables. However, this is computationally too expensive and also too hard to estimate accurately owing to the limited amount of training data. Therefore, we try to approximate the high-dimensional joint entropy term by a series of practically computable terms. We first rewrite the high-dimensional joint entropy term as a sum of a series of multivariate mutual information terms (a process we refer to as *Decompose*), and then we approximate it with a view to computational efficiency. The multivariate mutual information of a given variable set T can be defined by information theory (McGill, 1954):

$$I(\{T\}) = -\sum_{X\in T'}(-1)^{|X|}H(X) \tag{3}$$

Note that $\sum_{X\in T'}$ represents a sum over all elements X drawn from T', and T' denotes the power set of T. Suppose $T=\{f_1,f_2\}$, then $T'=\{\phi,f_1,f_2,\{f_1,f_2\}\}$. While the mutual information measures dependence between a pair of variables, multivariate mutual information can account for dependencies among multiple variables.

3.1. Decomposing high-dimensional entropy: $H(S)$ and $H(L)$

In this section, we decompose the high-dimensional joint entropy terms in Eq. (1): $H(S)$ and $H(L)$. Let S be a set of n features, and let X be one possible elements drawn from $S_k=\{e|e\in S,|e|=k\}$. Then, the sum of entropies over all elements whose cardinality is k can be defined as:

$$U_k(S) = \sum_{X\in S_k} H(X) \tag{4}$$

This notation is more useful when input set is a power set. Suppose $S=\{f_1,f_2,f_3\}$, then $U_2(S')=H(f_1,f_2)+H(f_1,f_3)+H(f_2,f_3)$ where $S'_k=\{e|e\in S',|e|=k\}$. Let Y be a possible element drawn from X'_m where $m\leqslant k\leqslant n$. By using Eq. (4), we obtain:

$$H(S) = \sum_{k=1}^{n}\sum_{m=1}^{k}(-1)^{k+m}\left(\sum_{X\in S'_k}\sum_{Y\in X'_m}H(Y)\right) \tag{5}$$

Proof is provided in Appendix A. To transform the high-dimensional joint entropy estimation into a series of k-dimensional joint entropy estimation problems, we decompose $H(S)$ into a sum of the pieces of multivariate mutual information using Eq. (5):

$$\begin{aligned} H(S) &= \sum_{k=1}^{n}\sum_{X\in S'_k}(-1)^k\left(\sum_{m=1}^{k}\sum_{Y\in X'_m}(-1)^m H(Y_m)\right) \\ &= -\sum_{X\in S'}(-1)^{|X|}\left(-\sum_{Y\in X'}(-1)^{|Y|}H(Y)\right) = -\sum_{X\in S'}(-1)^{|X|}I(\{X\}) \end{aligned} \tag{6}$$

Similar to Eq. (4), we now define the sum of multivariate mutual information with cardinality k as follows:

$$V_k(S) = \sum_{X\in S_k} I(\{X\}) \tag{7}$$

Then, we can rewrite Eq. (6) as follows:

$$H(S) = -\sum_{k=1}^{n}(-1)^k V_k(S') \tag{8}$$

As a result, $H(S)$ is represented by the right-hand side of Eq. (8), and the joint entropy of labels, $H(L)$, where $|L|=t$, can be represented as:

$$H(L) = -\sum_{k=1}^{t}(-1)^k V_k(L') \tag{9}$$

3.2. Decomposing the joint entropy of two sets: $H(S,L)$

We can decompose the joint entropy of the two sets using Eq. (5):

$$H(S,L) = -\sum_{k=1}^{n+t}(-1)^k V_k(\{S,L\}') \tag{10}$$

Detailed derivation is provided in Appendix B. Further, we can divide the power set $\{S,L\}'$ in the right hand side of Eq. (10) into three parts using power-set theorem; the first part has variable sets from $S'\times L'_0$, the second part has variable sets from $S'_0\times L'$, and third part is composed of remained subsets where $\times$ denotes the cartesian product of two sets. For example, $V_3(\{S,L\}')$ can be represented as a sum over these three parts; the first part is $V_3(S'_3\times L'_0)$, the second part is $V_3(S'_0\times L'_3)$, and the third part is $V_3(\{S'_2\times L'_1\})+V_3(\{S'_1\times L'_2\})$. Thus, $V_k(\{S,L\}')$ can be rewritten as:

$$V_k(\{S,L\}') = V_k(S'_k\times L'_0)+V_k(S'_0\times L'_k)+\sum_{p=1}^{k-1}V_p(S'_{k-p}\times L'_p) \tag{11}$$

Eq. (11) represents that any elements of cardinality k from $\{S,L\}'$ can be divided into elements from $S'_k=S'_k\times L'_0$, $L'_k=S'_0\times L'_k$, and $S'_{k-p}\times L'_p$. Because the third part represents multivariate mutual information among variables chosen from a combination of S and L, there are no terms with $k=1$. Hence, we rewrite Eq. (10) using Eq. (11) as follows:

$$\begin{aligned} H(S,L) &= -\sum_{k=1}^{n+t}(-1)^k\left(V_k(S'_k\times L'_0)+V_k(S'_0\times L'_k)\right) \\ &\quad -\sum_{k=2}^{n+t}(-1)^k\left(\sum_{p=1}^{k-1}V_p(S'_{k-p}\times L'_p)\right) \\ &= -\sum_{k=1}^{n}(-1)^k V_k(S') - \sum_{k=1}^{t}(-1)^k V_k(L') \\ &\quad -\sum_{k=2}^{n+t}\sum_{p=1}^{k-1}(-1)^k V_p(S'_{k-p}\times L'_p) \end{aligned} \tag{12}$$

We can rewrite the mutual information between two sets by combining Eqs. (8), (9), and (12):

$$\begin{aligned} I(S;L) &= H(S)+H(L)-H(S,L) \\ &= -\sum_{k=1}^{n}(-1)^k V_k(S') - \sum_{k=1}^{t}(-1)^k V_k(L') + \sum_{k=1}^{n}(-1)^k V_k(S') \\ &\quad +\sum_{k=1}^{t}(-1)^k V_k(L') + \sum_{k=2}^{n+t}\sum_{p=1}^{k-1}(-1)^k V_k(S'_{k-p}\times L'_p) \\ &= \sum_{k=2}^{n+t}\sum_{p=1}^{k-1}(-1)^k V_k(S'_{k-p}\times L'_p) \end{aligned} \tag{13}$$

Finally, we get a k-dimensional representation of the mutual information between two sets as shown in Eq. (13).

3.3. Feature selection algorithm

The present study focuses on developing a computationally efficient algorithm to select a compact set of features. Thus, for computational efficiency, we consider an approximated solution of Eq. (13) by constraining the calculations of $V_k(\cdot)$ functions with less than three cardinality. This can be obtained by replacing the limits $n+t$ in the summations of Eq. (13) with three. By rewriting $V_k(\cdot)$ functions into the multivariate mutual information terms, we obtain:

$$\begin{aligned}\widetilde{I}(S;L) &= V_2(S'_1 \times L'_1) - V_3(S'_2 \times L'_1) - V_3(S'_1 \times L'_2) \\ &= \sum_{f_i \in S}\sum_{l_j \in L} I(\{f_i, l_j\}) - \sum_{f_i \in S}\sum_{f_j \in S}\sum_{l_k \in L} I(\{f_i, f_j, l_k\}) \\ &\quad - \sum_{f_i \in S}\sum_{l_j \in L}\sum_{l_k \in L} I(\{f_i, l_j, l_k\}) \end{aligned} \tag{14}$$

It is worth noting that if we want to obtain a more accurate value of the mutual information between S and L, then we should calculate higher degree of relations, i.e., more than three, in Eq. (13); however, it is computationally expensive because calculations of multivariate mutual information becomes prohibitive. From Eq. (14), we can easily derive our feature selection algorithm in the circumstance of incremental selection. Suppose we have already selected a feature subset S; then we can see that to be selected feature f^+ should maximize $\widetilde{I}(\{S, f^+\}; L)$ in the incremental selection. Thus, the feature f^+ in each step should maximize the following equation:

$$\begin{aligned} J &= \widetilde{I}(\{S, f^+\}; L) - \widetilde{I}(S; L) \\ &= \sum_{f_i \in \{S f^+\}}\sum_{l_j \in L} I(\{f_i, l_j\}) - \sum_{f_i \in \{S f^+\}}\sum_{f_j \in \{S f^+\}}\sum_{l_k \in L} I(\{f_i, f_j, l_k\}) \\ &\quad - \sum_{f_i \in \{S f^+\}}\sum_{l_j \in L}\sum_{l_k \in L} I(\{f_i, l_j, l_k\}) - \sum_{f_i \in S}\sum_{l_j \in L} I(\{f_i, l_j\}) \\ &\quad + \sum_{f_i \in S}\sum_{f_j \in S}\sum_{l_k \in L} I(\{f_i, f_j, l_k\}) + \sum_{f_i \in S}\sum_{l_j \in L}\sum_{l_k \in L} I(\{f_i, l_j, l_k\}) \\ &= \sum_{l_i \in L} I(\{f^+, l_i\}) - \sum_{f_i \in S}\sum_{l_j \in L} I(\{f^+, f_i, l_j\}) - \sum_{l_i \in L}\sum_{l_j \in L} I(\{f^+, l_i, l_j\}) \end{aligned} \tag{15}$$

Algorithm 1. Proposed multi-label feature selection algorithm

1: **Input:** n; ▷ Number of to be selected features
2: **Output:** S; ▷ Selected feature subset
3: Initialize $S \leftarrow \{\phi\}$ and $F \leftarrow \{f_1, \ldots, f_d\}$;
4: **repeat**
5: Find the feature $f^+ \in F$ maximizing Eq. (15);
6: Set $S \leftarrow \{S \cup f^+\}$, and $F \leftarrow F \setminus S$;
7: **until** $|S| = n$
8: Output the set S containing the selected features;

Given a set of already selected features, the algorithm chooses the next features as the one that maximizes the criterion J under incremental selection. The proposed algorithm can be described by the Algorithm 1. Note that this is a greedy algorithm; even though it is not guaranteed to find the global maximum value of J, it provides a computationally efficient performance for various applications.

4. Experimental results

In this section, we verify the effectiveness of the proposed method by comparing its performance against conventional multi-label feature selection methods.

4.1. Data sets and evaluation

We experiment with three data sets from different applications; bioinformatics, semantic scene analysis, and text categorization. The biological data set, *Yeast*, is concerned with gene function classification of the Yeast *Saccharomyces cerevisiae* (Elisseeff and Weston, 2001). The gene functional classes made from the hierarchies that represent maximum 190 functions of genes. After preprocessing, top four levels of hierarchies were chosen to compose the multi-labeled functions. The image data set, *Scene*, is concerned with the semantic indexing of still scenes (Boutell et al., 2004). Each scene possibly contains multiple objects such as *desert*, *mountains*, *sea* and so on. Thus, those objects can be directly used to compose the multi-label of each still scene. The text data set, *Enron*, is a subset of the Enron email corpus (Klimt and Yang, 2004). An email may contain words according to its objective such as humor, admiration, friendship, and so on. Thus, words and objectives of an email can be naturally encoded into the multi-label data set. Table 1 displays certain standard statistics of the data sets such as the number of patterns, number of features, number of labels, and label density (Tsoumakas and Katakis, 2007).

Both the Yeast and Scene data sets consist of continuous features. We discretized the Yeast and Scene data sets using the Equal-width interval scheme to improve the computational efficiency (Dougherty et al., 1995); each continuous feature was then binarized into a categorical feature with two bins. Note that more complex discretization schemes can be applied, or we can directly calculate the multivariate mutual information from continuous features using multi-dimensional entropy estimation techniques (Beirlant et al., 1997; Miller, 2003; Lee, 2010).

We compared our proposed method to three conventional methods: ELA + χ^2, PPT + χ^2, and PPT + MI. The classification performance of the four feature selection methods, including the proposed method, was measured using the multi-label naive bayes (MLNB) classifier (Zhang et al., 2009). We evaluated the performance of the methods using a 30% hold-out set. Specifically, 70% of patterns randomly chosen from the data set were used for the training process, and the remaining 30% were used for measuring the performance of each feature selection method. Those experiments were applied 30 times iteratively, and the average value was taken to represent the classification performance. In the multi-label classification problem, performance can be assessed by several evaluation measures. We employed four conventional evaluation measures: Hamming loss, Ranking loss, Coverage, and multi-label accuracy (Boutell et al., 2004; Tsoumakas and Vlahavas, 2007). Let $P = \{(P_i, \lambda_i) | 1 \leqslant i \leqslant p\}$ be a given test set where $\lambda_i \subseteq L$ is a correct label subset, and $Y_i \subseteq L$ be a predicted label set corresponds to P_i. The Hamming loss is defined to be $hloss(P) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{l}|\lambda_i \triangle Y_i|$ where $\triangle$ denotes the symmetric difference between two sets. In most cases, a multi-label classifier could output the real-valued likelihood y_j between P_i and each label $l_j \in L$. The Ranking loss measures ranking quality of those likelihood, defined to as $rloss(P) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|\lambda_i||\overline{\lambda_i}|}\left|\{(y_1, y_2) | y_1 \leqslant y_2, (y_1, y_2) \in \lambda_i \times \overline{\lambda_i}\}\right|$ where $\overline{\lambda_i}$ denotes the complementary set of λ_i. Moreover, those likelihood can be ranked according to its value, for example, if $y_1 > y_2$ then $rank(y_1) < rank(y_2)$. Then the Coverage can be defined as $cov(P) = \frac{1}{p}\sum_{i=1}^{p}\max_{y \in \lambda_i} rank(y) - 1$. In addition, the multi-label

Table 1
Brief description of multi-label data sets.

Name	Domain	Patterns	Features	Labels	Density
Scene	Image	2407	294	6	0.179
Enron	Text	1702	1001	53	0.064
Yeast	Biology	2417	103	14	0.303

accuracy is defined to be $mlacc(P) = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{|\lambda_i \cap Y_i|}{|\lambda_i \cup Y_i|}\right)$. The Hamming loss evaluates how many times an pattern-label pair is misclassified, and other three measures concern the ranking quality of different labels for each test pattern. The first three evaluation measures indicate good classification performance when evaluated as low values, whereas the last evaluation measure, multi-label accuracy, indicates good classification performance when the classifier achieves high values. These four evaluation measures demonstrate different aspect of multi-label classification performance.

4.2. Comparison results

Fig. 1 shows the classification performance of each feature selection method for the *Scene* data set. The horizontal axis represents the size of the selected feature subset according to each feature selection method, and the vertical axis indicates the classification performance of certain evaluation measures. The proposed method showed superior classification performance to other conventional methods for any size of selected feature subset. Fig. 1(a) shows that the Hamming loss of the proposed method improved with the size of the selected feature subset. However, the Hamming loss of other conventional methods rapidly degraded as the size of the selected feature subset ranged from 1 to 10, and then, the performance slowly improved as the size of the feature subset grew larger. The Hamming loss of the feature subsets selected using ELA + χ^2, PPT + χ^2, and PPT + MI were 0.3344, 0.3356, and 0.3316, respectively, with 20 features selected, whereas the Hamming loss of the selected feature subset according to the proposed method was 0.1394. Thus, we can see that the performance of the proposed method showed an improvement of 0.1962 over the conventional PPT + χ^2. Fig. 1(b) shows that the Ranking loss was improved to a great extent by using the proposed method, and this tendency was consistent with the comparison results of the Coverage and multi-label accuracy, as shown in Fig. 1(c) and (d). The Ranking loss of the feature subsets selected using the proposed method was 0.1344 when 20 features were

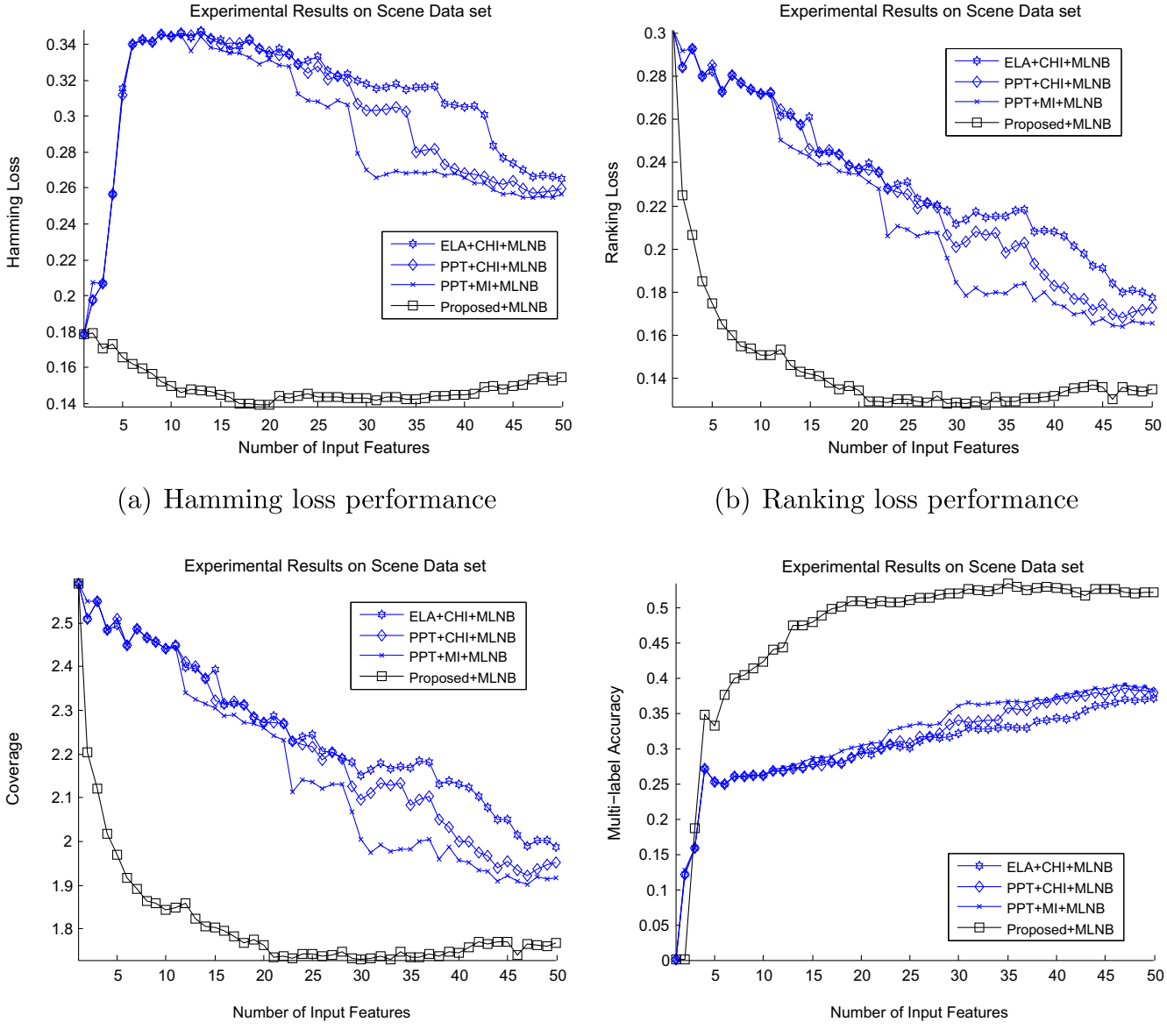

Fig. 1. Classification performance of the *Scene* data set according to feature subsets using the proposed method and three conventional feature selection methods: PPT + χ^2, PPT + MI, and ELA + χ^2.

selected, whereas the Ranking loss of the feature subsets selected using ELA + χ^2, PPT + χ^2, and PPT + MI were 0.2369, 0.2370, and 0.2344, respectively. Thus, we can conclude that the proposed method selected a more effective feature subset than the other conventional feature selection methods. The best classification performance evaluated by Hamming loss, Ranking loss, Coverage, and multi-label accuracy were all achieved using the proposed method, with scores of 0.1394, 0.1280, 1.7296, and 0.5348 with different sizes of selected feature subsets.

Fig. 2 shows the classification performance of each feature selection method for the *Enron* data set. The proposed method shows better classification performance compared to other conventional methods in terms of Hamming loss. Fig. 2(a) shows that our proposed method demonstrated superior performance compared to other conventional methods in the Hamming loss experiment. For any size of the selected feature subsets, the proposed method showed better performance than the other conventional methods. The Hamming loss of the feature subsets selected using ELA + χ^2, PPT + χ^2, and PPT + MI were 0.0976, 0.0930, and 0.0949, respectively, for 20 selected features, whereas the Hamming loss of the selected feature subset of the proposed method was 0.0631. Fig. 2(b) shows that the four feature selection methods started with a similar Ranking loss value until 15 features were selected. However, with the growing size of the selected feature subset, our proposed method showed better performance than three other methods. The Ranking loss of the feature subsets in accordance with the proposed method was 0.1251 when 20 features were selected, whereas ELA + χ^2, PPT + χ^2, and PPT + MI had achieved a Ranking loss of 0.1453, 0.1457, and 0.1400, respectively, for the same size of feature subsets. We can see a similar tendency in Fig. 2(c) and (d). Hence, our proposed method has shown its superiority compared to the other three conventional methods in the classification of the *Enron* data set.

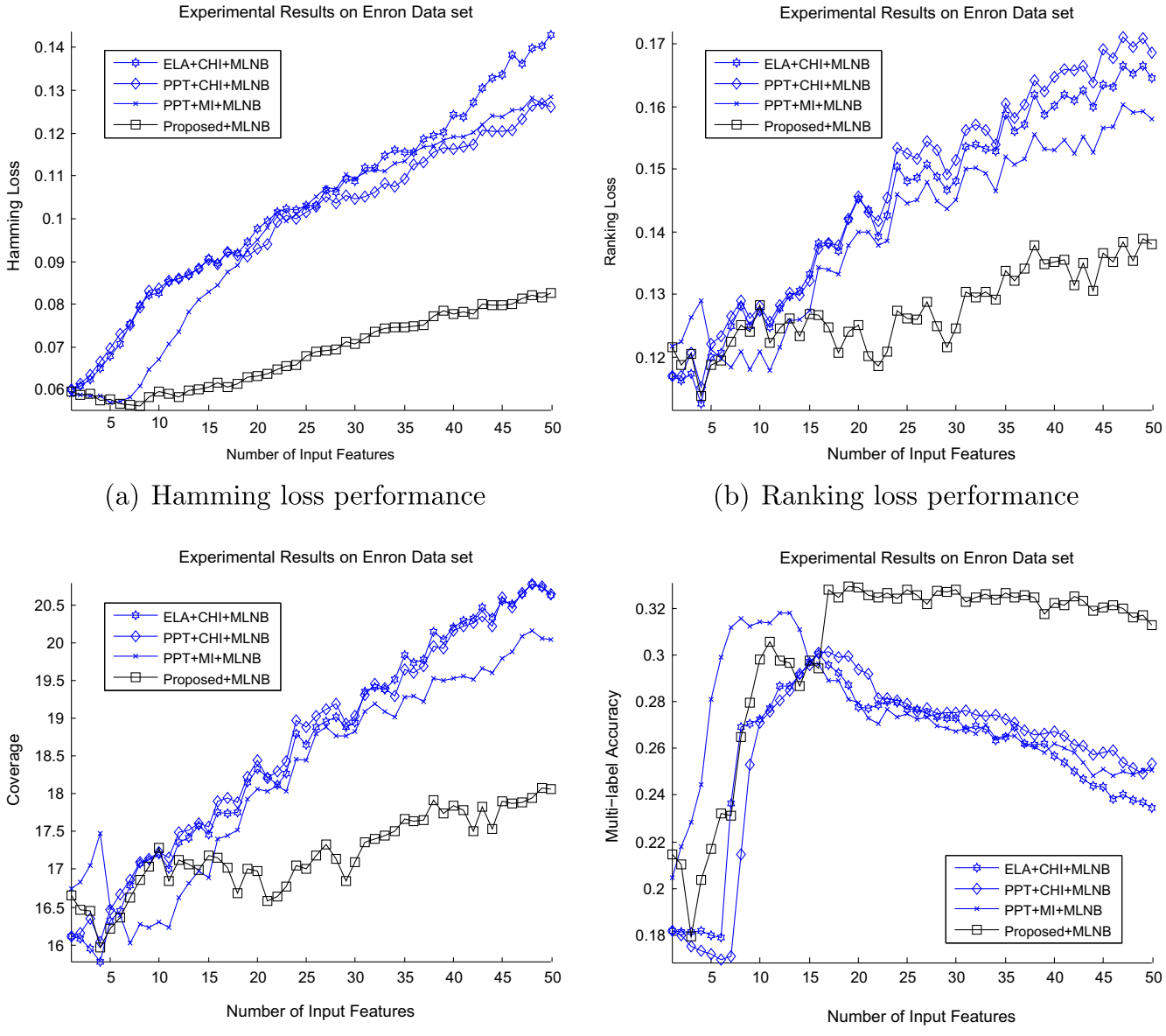

Fig. 2. Classification performance of the *Enron* data set according to feature subsets using the proposed method and three conventional feature selection methods: PPT + χ^2, PPT + MI, and ELA + χ^2.

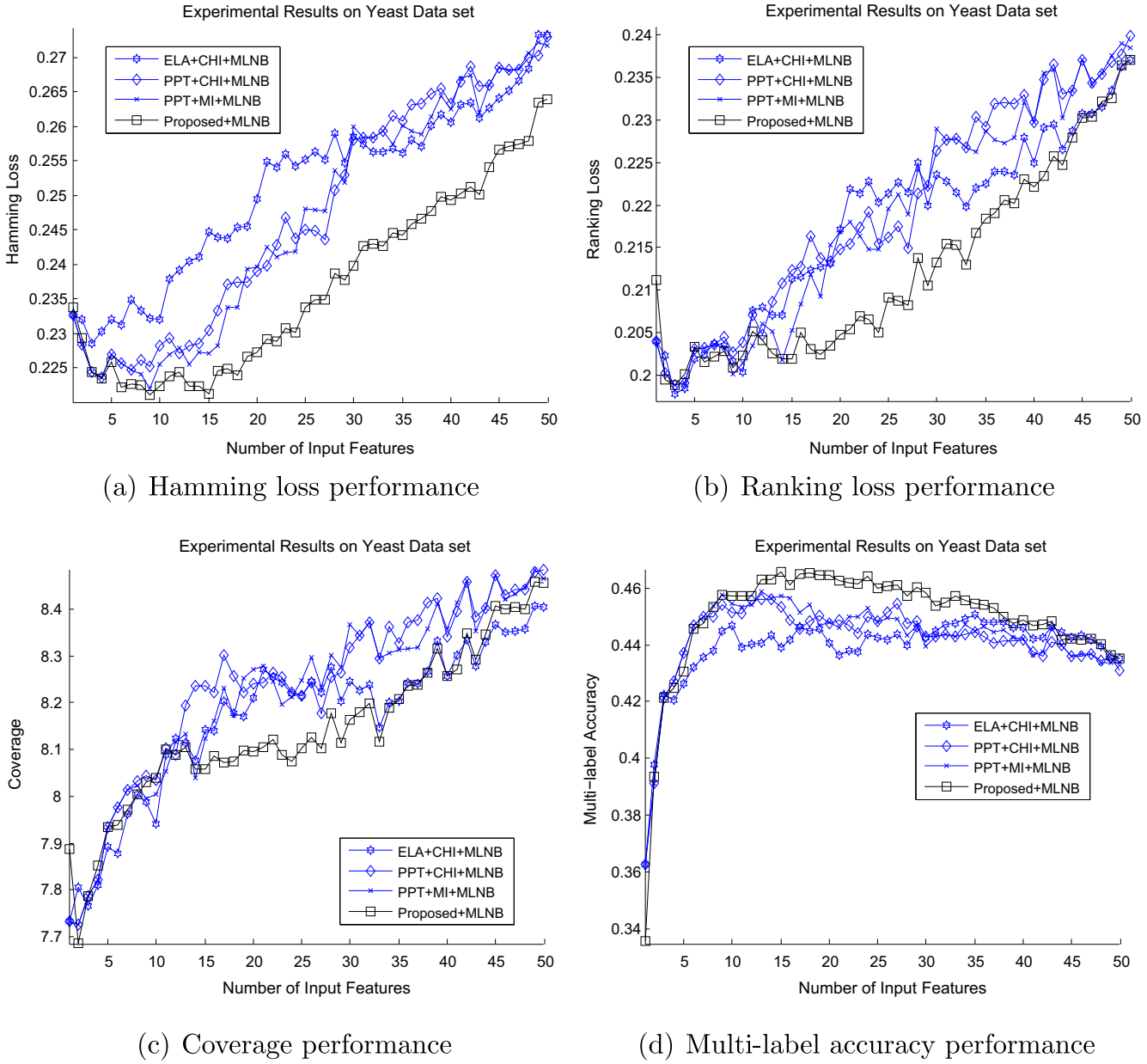

Fig. 3. Classification performance of the *Yeast* data set according to feature subsets using the proposed method and three conventional feature selection methods: PPT + χ^2, PPT + MI, and ELA + χ^2.

Fig. 3 shows the classification performance of each feature selection method for the *Yeast* data set. The comparison results for the Hamming loss are shown in Fig. 3(a). The proposed method always showed better Hamming loss compared to the other three conventional methods. The Hamming loss of the feature subsets selected using ELA + χ^2, PPT + χ^2, and PPT + MI were 0.2494, 0.2390, and 0.2397, respectively, for 20 selected features, whereas the Hamming loss of the selected feature subset of the proposed method was 0.2273. Thus, the proposed method shows better classification performance than other conventional methods in terms of Hamming loss. The Ranking loss of the proposed methods was shown to have better performance when the size of the selected features subset ranged from 15 to 30. The Ranking loss of the feature subsets in accordance with the proposed method was 0.2047 when 20 features were selected, whereas ELA + χ^2, PPT + χ^2, and PPT + MI had achieved 0.2171, 0.2148, and 0.2168, respectively, with the same size of feature subsets. The comparison results of both the Coverage and multi-label accuracy were similar to the Ranking loss results, as shown in Fig. 3(c) and (d). Although the proposed method showed better performance than other methods, the gain in classification performance was not large enough because the dependency of each feature in Yeast data set is similar to each other.

5. Conclusions

In this paper, we presented a multivariate mutual information-based feature selection method for multi-label classification. Our proposed method does not rely on any problem transformation method to select a relevant feature subset. To efficiently evaluate the dependency of input features in multivariate situations, the proposed method calculates three-dimensional interactions among features and labels instead of the calculating prohibitive high-dimensional density estimations. Our comprehensive experiments demonstrate that the

classification performance can be significantly improved by the proposed method. Comparison results on three real-world data sets emerging from different domains show the advantage of the proposed method compared with the three conventional methods based on three problem transformation methods and two score measures in terms of four multi-label performance measures: the Hamming loss, Ranking loss, Coverage, and multi-label accuracy. Thus, we showed that the proposed method can find very effective feature subsets for the multi-label classification problem.

Future work should include the study of the influence of the approximation accuracy; because the proposed method only considers three-dimensional interactions among features and labels, it may lose important information with higher-order label dependency. However, this leads to two practical difficulties: high-dimensional density estimation with limited size of training patterns and expensive computational cost of high-order multivariate mutual information. As a future work, we would like to study this issue further.

Appendix A

Eq. (5) indicates that the entropy of a variable set is equal to the sum of entropy of the power set of a set S with Möbius inversion. It is easy to imagine that if we consider the Pascal's Triangle, in which each element is determined by sum of two elements in its upper row. Suppose we take the values in a row except first row. In addition, if values located in even position at respective rows take positive, and values located in odd position take negative, then sum of these values should be zero. The proof of Eq. (5) uses this property; a sum of entropy of the power set of a set S with Möbius inversion would be equal to the entropy of S itself, since the entropy of power sets of S are self-eliminated by the binomial theorem except the entropy of the variable set S; a value located in the first row.

Let X be a possible elements drawn from S'_k, and let Y be a possible elements drawn from X'_m where $m \leqslant k \leqslant n$. Since our goal is to show the equality should be satisfied, we swap the left hand side of Eq. (5) and the right hand side for simplicity. Thus the equation is rewritten as:

$$\sum_{k=1}^{n}\sum_{m=1}^{k}(-1)^{k+m}\underbrace{\left(\sum_{X\in S'_k}\sum_{Y\in X'_m}H(Y)\right)}_{Part\ 1}=H(S) \tag{16}$$

Proof. Since *Part* 1 means that the sum over all possible elements from X'_k with the largest cardinality is constrained by k, we can see that this term can be represented as a sum of a series of several joint entropy terms. We illustrate the behavior of *Part* 1 with changing k and m in Table 2. For example, the series can be represented by Eq. (4) as $2U_1(S') = 2(H(f_1)+H(f_2)+H(f_3))$ when $k = 2$ and $m = 1$ in Table 2. Thus, we can see that *Part* 1 can be represented by a linear function of Eq. (4). Let $\sum_{X\in S'_k}\sum_{Y\in X'_m}H(Y) = \alpha_{k,m}U_m(S')$; then, Eq. (5) can be represented as:

$$\sum_{k=1}^{n}\sum_{m=1}^{k}(-1)^{k+m}\left(\sum_{X\in S'_k}\sum_{Y\in X'_m}H(Y)\right) = \sum_{k=1}^{n}\sum_{m=1}^{k}(-1)^{k+m}\alpha_{k,m}U_m(S') \quad \text{where } m\leqslant k \tag{17}$$

The coefficient $\alpha_{k,m}$ is determined by the generator. First, X is chosen from S'_k, and the number of possible $X \in S'_k$ can be represented as $\binom{n}{k}$. In addition, Y is generated from the element of X'_m with cardinality m, and thus, we simply represent it as $\binom{k}{m}$, where $1 \leqslant m \leqslant k$. Finally, the term $U_m(S')$ is composed of $\binom{n}{m}$ of the entropy terms. Thus, we can formalize the coefficient $\alpha_{k,m}$ as:

$$\alpha_{k,m} = \frac{\binom{n}{k}\binom{k}{m}}{\binom{n}{m}} = \frac{(n-m)!}{(n-k)!(k-m)!} \tag{18}$$

Further, we can simplify the $\alpha_{k,m}$ as $\binom{n-m}{k-m}$, where $m \leqslant k \leqslant n$. Thus, by combining Eqs. (17) and (18) using the Pascal's Triangle, we obtain:

$$\sum_{k=1}^{n}\sum_{m=1}^{k}(-1)^{k+m}\alpha_{k,m}U_m(S') = \sum_{k=1}^{n}\sum_{m=1}^{k}(-1)^{k+m}\binom{n-m}{k-m}U_m(S') \tag{19}$$

Eq. (19) represents the sum of entropies in kth row on Pascal's Triangle. We transform the limits of the summations by rewriting row-wise summations to column-wise summations in Pascal's Triangle as follows:

$$\sum_{k=1}^{n}\sum_{m=1}^{k}(-1)^{k+m}\binom{n-m}{k-m}U_m(S') = \sum_{m=1}^{n}\sum_{k=m}^{n}(-1)^{k+m}\binom{n-m}{k-m}U_m(S') \tag{20}$$

where $m \leqslant k$. Replace k with $k+m$:

$$\begin{aligned}&\sum_{m=1}^{n}\sum_{k=m}^{n}(-1)^{k+m}\binom{n-m}{k-m}U_m(S')\\ &=\sum_{m=1}^{n}\sum_{k+m=m}^{n-m}(-1)^{(k+m)+m}\binom{n-m}{(k+m)-m}U_m(S')\\ &=\sum_{m=1}^{n}\sum_{k=0}^{n-m}(-1)^{k}\binom{n-m}{k}U_m(S')\end{aligned} \tag{21}$$

Because $\sum_{j=0}^{n}(-1)^j\binom{n}{j} = 0$ where $n > 0$ by the binomial theorem, we separate $\sum_{m=1}^{n}$ into $\sum_{m=1}^{n-1}$ and $m = n$, since $m = n$ then $\sum_{k=0}^{n-m}$ turns to $\sum_{k=0}^{0}$ in Eq. (21), we obtain:

$$\begin{aligned}&\sum_{m=1}^{n}\sum_{k=0}^{n-m}(-1)^{k}\binom{n-m}{k}U_m(S') = \sum_{m=1}^{n-1}\sum_{k=0}^{n-m}(-1)^{k}\binom{n-m}{k}U_m(S') + U_n(S')\\ &=\sum_{m=1}^{n-1}\sum_{k=0}^{n-m}(-1)^{k}\binom{n-m}{k}U_m(S') + H(S) = \sum_{m=1}^{n-1}(0) + H(S) = H(S)\end{aligned} \tag{22}$$

Eq. (22) indicates that the entropy of variable set S can be written as a combination of the entropies computed from subsets. □

Table 2
An example of *Part* 1 in Eq. (5) when $S = \{f_1, f_2, f_3\}$. The X represents possible elements drawn from S'_k, whereas Y represents the possible elements drawn from X'_m. As seen in the table, the entropy sum according to k and m can be easily represented by Eq. (4).

	S'_1			S'_2			S'_3
	$\{f_1\}$	$\{f_2\}$	$\{f_3\}$	$\{f_1,f_2\}$	$\{f_1,f_3\}$	$\{f_2,f_3\}$	$\{f_1,f_2,f_3\}$
	$H(f_1)$	$H(f_2)$	$H(f_3)$	$H(f_1)$	$H(f_1)$	$H(f_2)$	$H(f_1)$
X'_1				$H(f_2)$	$H(f_3)$	$H(f_3)$	$H(f_2)$
							$H(f_3)$
				$H(f_1,f_2)$	$H(f_1,f_3)$	$H(f_2,f_3)$	$H(f_1,f_2)$
X'_2							$H(f_1,f_3)$
							$H(f_2,f_3)$
X'_3							$H(f_1,f_2,f_3)$

Appendix B

This relation can be easily confirmed using properties of the power set. Let S be a set of n variables, and S' denotes the power set of S. Suppose we add a set of t variables L. Then $\{S,L\}'$ can be represented as:

$$\{S,L\}' = \{S' \times L'\} = \{S'_0 \times L'_0, S'_0 \times L'_1, \ldots, S'_n \times L'_t\} \tag{23}$$

where $\times$ denotes the cartesian product between two sets. Let us illustrate a situation of including new variables to Eq. (8). Since a set of variables are newly included, we should consider additional relations among them. This can be written as:

$$H(S,L) = -\sum_{k=0}^{n}\sum_{m=0}^{t}(-1)^{k+m}V_{k+m}(S'_k \times L'_m) \tag{24}$$

Eq. (24) can be more simplified using $V_{k+m}(\cdot)$ function. For example, $V_{0+2}(S'_0 \times L'_2) + V_{1+1}(S'_1 \times L'_1) + V_{2+0}(S'_2 \times L'_0)$ can be represented as $V_2(S' \times L')$. Thus we can rewrite Eq. (24) as follows:

$$H(S,L) = -\sum_{k=0}^{n+t}(-1)^k V_k(S' \times L') \tag{25}$$

Since $V_0(\cdot) = 0$ and $\{S' \times L'\} = \{S,L\}'$ respectively, Eq. (25) can be rewritten as:

$$H(S,L) = -\sum_{k=1}^{n+t}(-1)^k V_k(\{S,L\}') \tag{26}$$

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0001772).

References

Beirlant, J., Dudewicz, E., Györfi, L., Van der Meulen, E., 1997. Nonparametric entropy estimation: An overview. Internat. J. Math. Statist. Sci. 6, 17–40.
Boutell, M., Luo, J., Shen, X., Brown, C., 2004. Learning multi-label scene classification. Pattern Recognition 37, 1757–1771.
Chen, W., Yan, J., Zhang, B., Chen, Z., Yang, Q., 2007. Document transformation for multi-label feature selection in text categorization. In: Proc. Seventh IEEE Internat Conf. of Data Mining (ICDM'07), pp. 451–456.
Diplaris, S., Tsoumakas, G., Mitkas, P., Vlahavas, I., 2005. Protein classification with multiple algorithms. Adv. Inf. 3746, 448–456.
Doquire, G., Verleysen, M., 2011. Feature selection for multi-label classification problems. Adv. Comput. Intell. 6691, 9–16.
Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: Internat. Worksh. Conf. on Machine Learning. Morgan Kaufmann Publishers, Inc., pp. 194–202.
Elisseeff, A., Weston, J., 2001. A kernel method for multi-labelled classification. Adv. Neural Inf. Process. Systems 14, 681–687.
Gu, Q., Li, Z., Han, J., 2011. Correlated multi-label feature selection. In: Proc. 20th ACM Internat. Conf. on Information and Knowledge Management. ACM, pp. 1087–1096.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Machine Learn. Res. 3, 1157–1182.
Klimt, B., Yang, Y., 2004. The enron corpus: A new dataset for email classification research. Lect. Note Comput. Sci. 3201, 217–226.
Lee, I., 2010. Sample-spacings-based density and entropy estimators for spherically invariant multidimensional data. Neural Comput. 22, 2208–2227.
Lewis, D., Yang, Y., Rose, T., Li, F., 2004. Rcv1: A new benchmark collection for text categorization research. J. Machine Learn. Res. 5, 361–397.
McGill, W., 1954. Multivariate information transmission. IRE Trans. Inf. Theory 4, 93–111.
Miller, E., 2003. A new class of entropy estimators for multi-dimensional densities. In: Proc. 2003 IEEE Internat. Conf. on Acoustic, Speech and Signal Processing (ICASSP'03). IEEE, pp. 297–300.
Read, J., 2008. A pruned problem transformation method for multi-label classification. In: Proc. 2008 New Zealand Comput. Sci. Res. Stud. Conf. (NZCSRS'08), pp. 143–150.
Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517.
Schapire, R., Singer, Y., 2000. Boostexter: A boosting-based system for text categorization. Machine Learn. 39, 135–168.
Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1–47.
Sun, Y., Wong, A., Kamel, M., 2009. Classification of imbalanced data: A review. Int. J. Pattern Recognition Artif. Intell. 23, 687.
Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I., 2008. Multilabel classification of music into emotions. In: Proc. Ninth Internat. Conf. Music Inform. Retrieval (ISMIR'08), Philadelphia, PA, USA.
Tsoumakas, G., Katakis, I., 2007. Multi-label classification: An overview. Internat. J. Data Warehouse Min. 3, 1–13.
Tsoumakas, G., Katakis, I., Vlahavas, I., 2011. Random k-labelsets for multi-label classification. IEEE Trans. Knowl. Data Eng. 23, 1079–1089.
Tsoumakas, G., Vlahavas, I., 2007. Random k-labelsets: An ensemble method for multilabel classification. Machine Learn. (ECML'07) 4701, 406–417.
Watanabe, S., 1969. Knowing and Guessing: A Quantitative Study of Inference and Information. Wiley, New York.
Yang, Y., Pedersen, J., 1997. A comparative study on feature selection in text categorization. In: Proc. 14th Internat. Conf. on Machine Learning, pp. 412–420.
Zhang, M., Peña, J., Robles, V., 2009. Feature selection for multi-label naive bayes classification. Inf. Sci. 179, 3218–3229.
Zhang, M., Zhou, Z., 2007. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40, 2038–2048.