LETTER

# Multi-Label Learning Using Mathematical Programming

**Hyunki LIM**[†], **Jaesung LEE**[†], ***Nonmembers***, ***and*** **Dae-Won KIM**[†a)], ***Member***

**SUMMARY** We propose a new multi-label feature selection method that does not require the multi-label problem to be transformed into a single-label problem. Using quadratic programming, the proposed multi-label feature selection algorithm provides markedly better learning performance than conventional methods.

***key words:*** *multi-label learning, feature selection, quadratic programming*

## 1. Introudction

Multi-label learning, i.e., the task of assigning an object to multiple categories simultaneously, has emerged as a popular problem for text categorization, multiple music-emotion recognition, and semantic scene annotation [1]–[3]. In multi-label learning, the accuracy of the output is strongly influenced by the input features. Hence there is a strong need for feature selection techniques that enhance multi-label learning [3]–[5]. In practical terms, multi-label feature selection methods must consider multiple labels concurrently; that is, an objective function should be able to evaluate the significance of a selected feature subset from the viewpoint of multiple labels. Previous studies, however, used a pre-processing method that transforms the multi-label problem into a single-label multi-class problem and then applied single-label feature selection methods [5]. Such problem transformation-based methods may cause information about the relationships among labels to be lost. In addition, they do not capture the relationships among features, but the relationships between a feature and a transformed label [2], [3]. In this study, we propose a multi-label feature selection method based on a novel objective function that is formulated as a quadratic programming (QP) problem using information theory. The experimental results show that the proposed method provides significant benefits over conventional problem transformation-based multi-label feature selection methods. To the best of our knowledge, this is the first multi-label feature selection method based on mathematical programming.

## 2. Related Work

Many studies have proposed two-step algorithms for multi-label feature selection. In general, the first step transforms the multi-label problem into a single-label problem, and the second step determines the importance of each feature in accordance with the transformed label set. These measures evaluate the effectiveness of each feature and select high-rank features. Trohidis et al. proposed a transform method named Label Powerset (LP), and used the $\chi^2$ statistic (CHI) as a scoring function for the retrieval of music information, specifically the recognition of six emotions that are simultaneously evoked by music clips [3]. LP transforms a multi-label to a single label by assigning each pattern's label set to a single class. Although LP directly considers label dependency, it yields considerably many classes. As a result, there are considerably few patterns assigned to each class. Chen et al. proposed Entropy-based Label Assignment (ELA), which also use CHI as a scoring function [8]. ELA assigns weights to a multi-labeled pattern based on the label entropy. It was argued that the learning algorithm could avoid overfitting, but ELA tends to lose information regarding the dependency among labels. Read proposed a Pruned Problem Transformation (PPT) that improved the LP [9]. In the training process, this method removes patterns with infrequent labels according to a predefined threshold $\tau$. However, the accuracy of multi-label learning may be limited if an inappropriate value of $\tau$ is selected. Conventional methods usually apply the ReliefF (RF) algorithm [5] or CHI to feature selection after transforming the multi-label to a single label.

Recently, Rodriguez-Lujan et al. proposed a Quadratic Programming Feature Selection (QPFS) method [10]. QPFS formulates a quadratic objective function for the single-label feature selection problem. The goal of the objective function is to maximize the dependency between the features and the label, and minimize the dependency among features to avoid redundancy. While considering two objectives concurrently, QPFS assigns a weight to each feature to satisfy the objective function. Thus, QPFS is not a heuristic greedy approach, but has a mathematical foundation and draws on a wider perspective. However, this method considers only single-label datasets. In this paper, we propose the first quadratic programming feature selection method for multi-label datasets. We formulate the relations among features and labels as a quadratic objective function. As a result, a QP solver naturally finds the weight of each feature by solving the given objective function.

## 3. Proposed Method

Our goal is to formulate an objective function that can be

Manuscript received July 9, 2014.
Manuscript publicized September 29, 2014.

[†]The authors are with School of Computer Science and Engineering, Chung-Ang University, Seoul 156–756, Korea.
a) E-mail: dwkim@cau.ac.kr

DOI: 10.1587/transinf.2014EDL8139

solved by a QP solver. To simultaneously consider (1) the dependency among features, and (2) the dependency between features and labels, two concepts should be incorporated in one objective function. Moreover, the objective function should consider the importance of features to perform multi-label feature selection.

Given $N$ features, input $F = \{f_1, \ldots, f_N\}$ and the label set $Y = \{y_1, \ldots, y_M\}$, multi-label feature selection aims to find a feature subset $S \subset F$ with $n \ll N$ features. The proposed method solves this problem by 1) finding an $N$-dimensional vector $x \in \mathbb{R}^N$ that contains suitable feature weights; and 2) selecting the $n$ features with the highest weight values. Because the number of features being selected is limited to $n$, similar features should not be included in $S$ concurrently. Thus, dependency among the selected features in $S$ should be minimized, whereas dependency between $S$ and $Y$ should be maximized. This concept can be naturally represented in the QP objective function. Our goal is to find a weight vector $x$ that minimizes the given objective function $f(x)$, written as

$$f(x) = \frac{1}{2} x^T Q x - c^T x \tag{1}$$

subject to $x_1, \ldots, x_N \geq 0$.

Let the symmetric positive semidefinite matrix $Q \in \mathbb{R}^{N \times N}$ represent the dependency among the features of $F$. In this work, $Q$ is computed using the total dependency of $F$ [6], written as

$$\begin{aligned} C(F) &= \sum_{i=1}^{N} H(f_i) - H(F) = \sum_{i=1}^{N} H(f_i) - H(f_1, \ldots, f_N) \\ &= \sum_{f_i, f_j \in F} I(f_i, f_j) - \sum_{f_i, f_j, f_k \in F} I(f_i, f_j, f_k) + \cdots \\ &\quad + (-1)^N I(f_1, \ldots, f_N) \text{ for } i \neq j \neq k \end{aligned} \tag{2}$$

where $I(T) = \sum_{U \subseteq T} (-1)^{|U|} H(U)$ is the interaction information of a feature subset $T$, and $H(T) = -\sum_{t \in T} P(t) \log P(t)$ is the joint entropy of $T$. Because the computational cost of calculating $C(F)$ increases exponentially with $N$, and $N$ is typically a large value in feature selection problems, this is computationally prohibitive. To circumvent this, we relax the computational cost of $C(F)$ by taking the first-order interaction information of $F$ because $Q$ is naturally able to represent the dependency between pairs of features:

$$Q_{ij} = I(f_i, f_j) \tag{3}$$

where $Q_{ij} \in Q$ represents the dependency between $f_i$ and $f_j$.

A non-negative vector $c \in \mathbb{R}^N$ in (1) represents the dependency between a feature $f_i$ and the multiple labels in the set $Y$; This can be computed using mutual information:

$$I(f_i; Y) = H(f_i) + H(Y) - H(f_i, Y) \tag{4}$$

Because $Y$ is a set of labels, the number of joint states in $Y$ increases exponentially according to the size of $Y$. Therefore, the calculation of $H(Y)$ becomes prohibitive when $M$ is a large value. Using total dependency, (4) can be rewritten as

$$\begin{aligned} I(f_i; Y) &= H(f_i) + H(Y) - H(f_i, Y) \\ &= H(f_i) + \left( \sum_{k=1}^{M} H(y_i) - C(Y) \right) - \\ &\quad \left( H(f_i) + \sum_{k=1}^{M} H(y_i) - C(f_i, Y) \right) \\ &= C(f_i, Y) - C(Y) \end{aligned} \tag{5}$$

As shown in (2), total dependency can be rewritten using the interaction information over all possible subsets of the given variables. Because $C(f_i, Y)$ and $C(Y)$ share variable subsets, except for new subsets brought about by $f_i$, the interaction information terms from $Y$ only in $C(f_i, Y)$ are cancelled by $C(Y)$. Therefore, (5) can be rewritten as

$$\begin{aligned} C(f_i, Y) - C(Y) = &\sum_{y_j \in Y} I(f_i, y_j) - \sum_{y_j, y_k \in Y} I(f_i, y_j, y_k) \\ &\cdots + (-1)^{M+1} I(f_i, y_1, \ldots, y_M) \end{aligned} \tag{6}$$

Similar to our formulation of $Q$ in (3), we relax (6) by taking the first-order interaction information between $f_i$ and $Y$. As a result, $c$ can be computed as

$$c_i = \sum_{y_j \in Y} I(f_i, y_j) \tag{7}$$

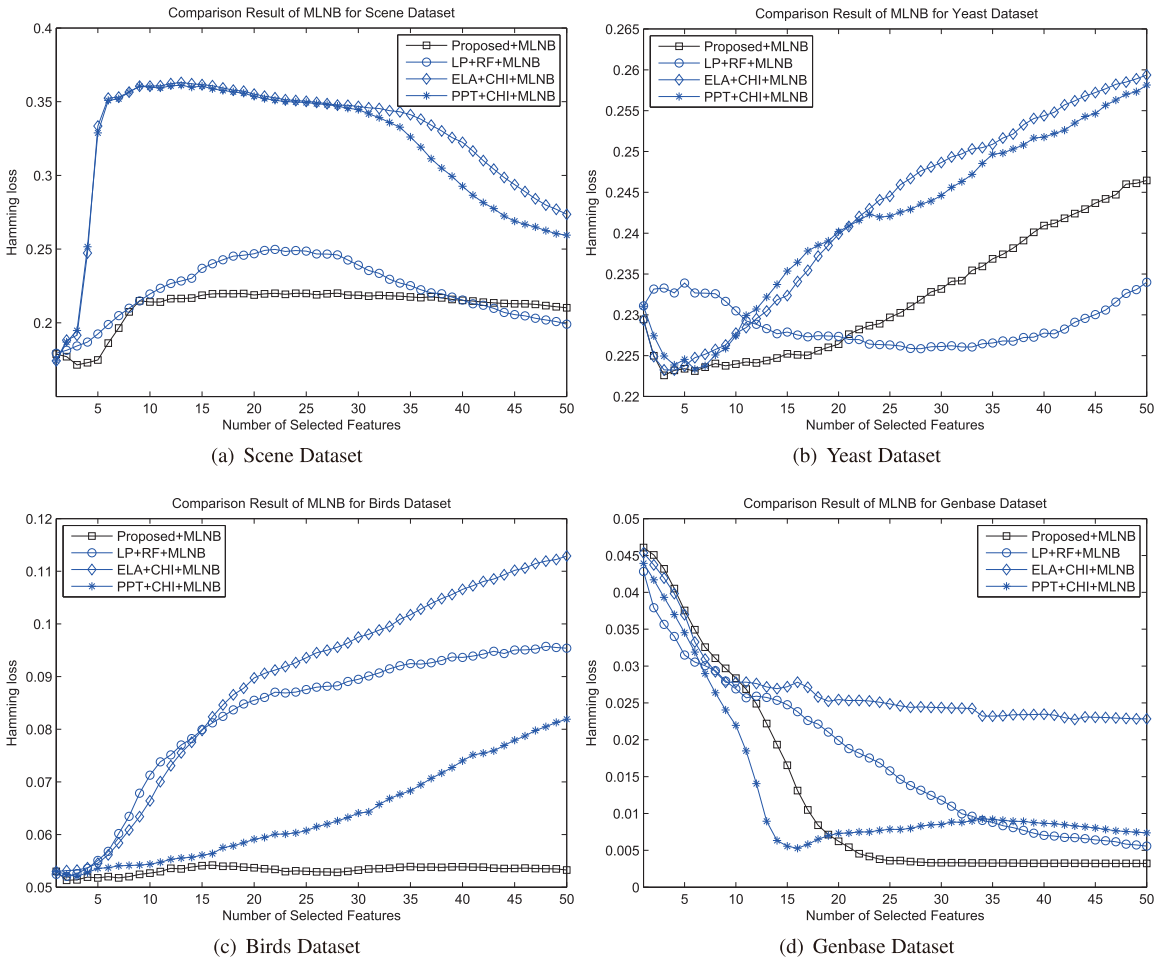
After minimizing (1) for the given $Q$ and $c$, the elements of $x$ represent the weight of each feature. Therefore, the selected feature subset $S$ can be obtained by including the $n$ features with the highest weight values $x_i$. Algorithm 1 outlines the procedure of the proposed method.

**Algorithm 1** Procedures of proposed method

```
1: procedure PROPOSED METHOD(F, Y, n)
2:     initialize c using (7) and Q using (3)        ▷ Initialization
3:     solve arg min f(x) of (1) using a QP solver   ▷ Obatin x
4:     sort F according to weights x in descending order
5:     S ← top n features in F                       ▷ Obtain output feature subset S
6: end procedure
```

## 4. Experimental Results

We compared the proposed method with three transformation-based multi-label feature selection methods [5]. LP + RF, ELA + CHI, and PPT + CHI. For the proposed method, we employed the active-set method as the QP solver. The feature subsets selected by each multi-label feature selection method were evaluated using a Multi-Label Naive Bayes (MLNB) classifier [4]. Table 1 lists the datasets [7] used in our experiments; these have been widely used for comparative purposes in multi-label classification. The performance was assessed using four evaluation measures: Hamming loss, Ranking loss, Coverage, and Subset accuracy [1], [4]. Low values of the Hamming loss, Ranking loss, and Coverage, and high values of Subset accuracy, indicate good multi-label classification performance.




(a) Scene Dataset (b) Yeast Dataset

(c) Birds Dataset (d) Genbase Dataset

**Fig. 1** Performance comparison of the proposed method and conventional feature selection methods.

**Table 1** Datasets used in the experiments.

| Datasets | Patterns | Features | Labels |
|---|---|---|---|
| Scene | 2,407 | 294 | 6 |
| Yeast | 2,417 | 103 | 14 |
| Birds | 645 | 260 | 19 |
| Genbase | 662 | 1,185 | 27 |

**Table 2** Performance comparison for each method.

| Measures | Datasets | Proposed | LP+RF | ELA+CHI | PPT+CHI |
|---|---|---|---|---|---|
| Hamming loss | Scene | **0.171**† | 0.179 | 0.174 | 0.174 |
| | Yeast | **0.223** | 0.226 | 0.223 | 0.223 |
| | Birds | **0.051**† | 0.052 | 0.053 | 0.052 |
| | Genbase | **0.003**† | 0.006 | 0.023 | 0.005 |
| Ranking loss | Scene | **0.118**† | 0.159 | 0.175 | 0.163 |
| | Yeast | **0.193** | 0.194 | 0.194 | 0.200 |
| | Birds | **0.080** | 0.092 | **0.080** | **0.080** |
| | Genbase | 0.007 | 0.006 | 0.044 | **0.006** |
| Coverage | Scene | **1.676**† | 1.893 | 1.924 | 1.873 |
| | Yeast | **7.616**† | 7.701 | 7.666 | 7.720 |
| | Birds | **2.813** | 3.159 | **2.813** | **2.813** |
| | Genbase | 1.558 | 1.579 | 2.609 | **1.557** |
| Subset accuracy | Scene | 0.196 | **0.251**† | 0.100 | 0.121 |
| | Yeast | **0.136**† | 0.128 | 0.129 | 0.127 |
| | Birds | **0.488** | 0.472 | 0.449 | 0.486 |
| | Genbase | **0.929**† | 0.895 | 0.533 | 0.883 |

† indicates that the proposed method is statistically superior to the conventional methods based on the paired $t$-test (0.05 significance level).

Figure 1 shows the Hamming loss performance of each method. In each figure, the horizontal axis represents the size of the selected feature subset and the vertical axis represents the Hamming loss value of the selected feature subset. In Fig. 1 (a), we can see that the proposed method outperforms ELA+CHI and PPT+CHI regardless of the feature subset size. For the Scene dataset, the proposed method achieves the best Hamming loss performance when the number of input features is about 5. Figure 1 (b) shows the Hamming loss performance for the Yeast dataset. In this figure, the Hamming loss performance of the four methods is similar. The proposed method, ELA+CHI, and PPT+CHI achieve optimal performance with a small number of input features, unlike LP+RF. However, as the number of feature subsets increases, the Hamming loss performance of ELA+CHI and PPT+CHI becomes much worse than that of the proposed method. In Fig. 1 (c), it can clearly be seen that the proposed method outperforms the other three methods for all the numbers of input features. The proposed method gives consistently low Hamming loss values as the size of the selected feature subset increases, whereas the Hamming

loss of the other three methods increases continuously. The results of the Genbase dataset, shown in Fig. 1 (d), show that the proposed method achieves optimal the best performance with 30 input features. The other three methods cannot match the Hamming loss performance obtained by the proposed method.

Table 2 summarizes the performance of each multi-label feature selection method for each dataset. The best value obtained for each dataset is marked in bold. Although conventional multi-label feature selection methods sometimes give better performance for three of the evaluation measures, the experimental results indicate that the proposed method gave the best performance in most experiments. This was confirmed by a paired $t$-test (0.05 significance level, denoted by "†" in the table).

## 5. Conclusion

We proposed a QP multi-label feature selection method based on information theory. An effective feature subset for multi-label learning was determined using the QP framework without resorting to problem transformation methods. To efficiently calculate the dependency of a dataset, we used first-order interaction information. Our comprehensive experiments demonstrated the improvement in classification performance produced by the proposed method.

Future work will focus on decreasing the time complexity of calculating the $Q$ matrix. Initializing of $Q$ for a given objective function may consume significant computational resources if $F$ is composed of many features. In this case, a matrix approximation technique or parallel programming can be used.

## Acknowledgements

This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2014.

### References

[1] M. Zhang and Z. Zhou, "A review on multi-label learning algorithm," Proc. IEEE Trans. Knowl. Data Eng., vol.99, p.1, March 2013.
[2] J. Lee and D.-W. Kim, "Feature selection for multi-label classification using multivariate mutual information," Pattern Recognit. Lett., vol.34, pp.349–357, Feb. 2013.
[3] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music into emotions," Proc. 9th Int. Conf. Music Information Retrieval, vol.8, pp.325–330, Philadelphia, USA, Sept. 2008.
[4] M. 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.
[5] 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, no.1, pp.135–151, 2013.
[6] S. Watanabe, "Information theoretical analysis of multivariate correlation," IBM J. Res. Devel., vol.4, no.1, pp.66–82, 1960.
[7] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, "MULAN: A java library for multi-label learning," J. Mach. Learn. Res., vol.12, no.7, pp.2411–2414, 2011.
[8] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang, "Document transformation for multi-label feature selection in text categorization," Seventh IEEE International Conference, pp.451–456, 2007.
[9] J. Read, "A pruned problem transformation method for multi-label classification," Proc. New Zealand Computer Science Research Student Conference, pp.143–150, 2008.
[10] I. Rodriguez-Lujan, R. Huerta, C. Elkan, and C.S. Cruz, "Quadratic programming feature selection," J. Mach. Learn. Res., vol.11, pp.1491–1516, 2010.