Low-Rank Approximation for Multi-label Feature Selection

Hyunki Lim, Jaesung Lee, and Dae-Won Kim

***Abstract*—The goal of multi-label feature selection is to find a feature subset that is dependent to multiple labels while maintaining as small number of features as possible. To select a compact feature subset, feature selection approaches that considers the dependency among features during its multi-label feature selection process. However, multi-label feature selection methods considering feature dependency suffer from its time-consuming task because the process of considering dependency among features consumes additional computational cost. In this paper, we propose a fast multi-label feature selection method considering feature dependency. The proposed method circumvents the prohibitive computations originated from the calculation of feature dependency by using an approximation. Empirical results conducted on several multi-label datasets demonstrate that the proposed method outperforms recent multi-label feature selection methods in terms of execution time.**

***Index Terms*—feature dependency, multi-label feature selection, mutual information, quadratic programming.**

I. Introduction

Recently with the advancement of multi-label data analysis [1], [2], the researches for knowledge mining on modern application areas give precious knowledge for achieving distinctive objectives of corresponding area. Such application areas include conventional text categorization [3], [4], image annotation which an image contains multiple objects [5], music analysis through acoustic information of music clips that expresses multiple emotions simultaneously [6], sentiment analysis for brand and social network service [7], and so on [8], [9]. A practical limitation can be caused if given multi-label dataset is composed of large number of features. This degrades the learning speed of machine learning algorithms, the generality of the knowledge, and the interpretability of explored model [10], [11]. The multi-label feature selection is considered an effective solution for achieving this limitation [12], [13].

Conventional multi-label feature selection methods evaluate the importance of each feature independently, thereby the dependency among features are ignored [14], [15]. As a result, a compact multi-label feature subset cannot be obtained because the selected feature subset contains redundant features, i.e. features similar to each other, if original multi-label dataset is composed of many redundant features [13]. To achieve this practical problem, a multi-label feature selection method must consider the feature dependency during its feature selection process. However, these methods commonly suffer from additional computational cost for evaluating feature dependency. To circumvent computational cost of feature dependency, Nyström method was proposed which is one of low-rank approximation methods [16]. Low-rank approximation methods are widely used for matrix approximation areas [17], [18]. Nyström method assumes that the matrix is kernel matrix. However, feature dependency does not meet the assumption, thus applying the low-rank approximation for the feature dependency cannot be appropriate.

In this paper, we propose a multi-label feature selection method by accelerating the process of evaluating the feature dependency. We design the quadratic function for evaluating the feature set and approximate the feature dependency by using heuristic method.

II. Procedure for Paper Submission

Let $W \in \mathbb{R}^N$ denote an input space constructed from a set of features F, where $|F| = N$ and patterns drawn from W are assigned to a certain label subset $\lambda \subseteq Y$, where $Y = \{y_1, \cdots, y_M\}$ is a finite set of labels with $|Y| = M$. The feature selection problem is to select a subset S composed of selected n features from $F(n \ll N)$, which jointly have the largest dependency on multiple labels Y.

A. Objective Function Modeling

We formulated an objective function that simultaneously considers the dependency among features, and the dependency between features and labels in previous study [19]. The proposed method solves the problem that minimize the objective function by finding an N-dimensional vector $x \in R^N$ that contains suitable feature weights; and select the n features with the highest weight values. Because the number of features being selected is limited ton, 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 quadratic function. Our goal is to find a weight vector x that minimizes the given objective function $f(x)$, written as

$$f(x) = \frac{1}{2}x^TQx - c^Tx \quad (1)$$

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

In this work, Q is computed using the mutual information as:

$$Q_{ij} = I(f_i, f_j) \quad (2)$$

Manuscript received October 30, 2015; revised January 13, 2016. 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 2015.

The authors are with the Department of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea (e-mail: hyunki05@gmail.com, jslee.cau@gmail.com, dwkim@cau.ac.kr).

doi: 10.18178/ijmlc.2016.6.1.569

where $Q_{ij} \in Q$ represents the dependency between f_i and f_j. $I(f_i, f_j)$ is calculated as:

$$I(f_i; f_j) = H(f_i) + H(f_j) - H(f_i, f_j) \tag{3}$$

where $H(T) = -\sum_{t \in T} p(t)\, log\, p(t)$. The vector c of (1) is calculated as:

$$c_i = \sum_{y_j \in Y} I(f_i; y_j) \tag{4}$$

where c_i represents dependency between f_i and labels. Detailed information of (1) is presented in [19].

B. Approximating Feature Dependency

In this section, we write I_{ij} and H_i in place of $I(f_i; f_j)$ and $H(f_i)$ for the space issue and readability. Because the computational cost for obtaining Q increases exponentially with N, and N is prevalently a large value in feature selection problems, this is computationally prohibitive. Then we propose simple heuristic method.

We can represent matrix Q as a block matrix

$$Q = \begin{pmatrix} A & B \\ B^T & E \end{pmatrix} \tag{5}$$

where $A \in \mathbb{R}^{k \times k}$, $B \in \mathbb{R}^{k \times (N-k)}$, and $E \in \mathbb{R}^{(N-k) \times (N-k)}$. Suppose we only know $[A\, B]$ of matrix Q. Then we approximate Q_{pq} in block matrix E.

$$Q_{pq} \approx \frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^{k} Q_{pi} + \frac{1}{k}\sum_{i=1}^{k} Q_{iq}\right) \tag{6}$$

The proposed approximation method means the average of the feature dependencies including index p or q from $[A\, B]$.

To show the superiority of the proposed method, we compare the Nyström method through an example that k is 1 in (5). Let $Q \in \mathbb{R}^{n \times n}$ be a symmetric matrix contained feature dependencies. We can represent matrix Q as a block matrix like (5) where $A \in \mathbb{R}$, $B \in \mathbb{R}^{1 \times (N-1)}$, and $E \in \mathbb{R}^{(N-1) \times (N-1)}$. Suppose we only know row vector $[A\, B]$ of matrix Q.

In submatrix E of feature dependency matrix Q based on MI, one element Q_{ij} can be approximated by the proposed method.

Proposition 1: When we know only $[A\, B]$ of matrix Q in Eq. (5) where $A \in \mathbb{R}$ and $B \in \mathbb{R}^{1 \times (N-1)}$, the proposed method approximates one element Q_{ij} of feature dependency matrix Q using

$$Q_{ij} \approx \frac{I_{1i} + I_{1j}}{2} \tag{7}$$

Then we can define the error of the proposed method about one element Q_{ij} for feature dependency approximation.

Lemma 1: When we know only $[A\, B]$ of matrix Q in Eq. (5) where $A \in \mathbb{R}$ and $B \in \mathbb{R}^{1 \times (N-1)}$, the approximating error E_{pro} of one element Q_{ij} of the proposed for feature dependency can be defined as

$$E_{pro} = \left| I_{ij} - \frac{I_{1i} + I_{1j}}{2} \right| \tag{8}$$

Theorem 1: When the (9) is satisfied, the error of approximation of the Nyström method is always bigger than the error of the proposed method or same.

$$8I_{ij} \geq 3(I_{1i} + I_{1j}) \tag{9}$$

The approximating error E_{Nys} of one element Q_{ij} of the Nyström method for feature dependency is defined in APPENDIX.

Proof. To show the difference of two errors, we can write the expression as the subtraction of squares of two errors and the multiplication of two terms.

$$E_{Nys}^2 - E_{Pro}^2 = (E_{Nys} - E_{Pro}) \times (E_{Nys} + E_{Pro}) \tag{10}$$

We can derive the left term of (10) as

$$\begin{aligned}(E_{Nys} - E_{Pro}) &= (I_{ij} - \frac{I_{1i}I_{1j}}{2H_1}) - (I_{ij} - \frac{I_{1i} + I_{1j}}{2}) \\ &= \frac{I_{1i}+I_{1j}}{2} - \frac{I_{1i}I_{1j}}{2H_1} \\ &= \frac{1}{2H_1}(H_1 I_{1i} + H_1 I_{1j} - I_{1i}I_{1j})\end{aligned} \tag{11}$$

The left term of (10) is always greater than or equal to 0 because $H_1 \geq I_{1i}$ and $H_1 \geq I_{1j}$ in information theory. In the same way, we can derive right term of (10) as

$$\begin{aligned}(E_{Nys} + E_{Pro}) &= 2I_{ij} - \frac{I_{1i}I_{1j}}{2H_1} - \frac{I_{1i} + I_{1j}}{2} \\ &= \frac{1}{2H_1}(4H_1 I_{ij} - I_{1i}I_{1j} - H_1 I_{1i} - H_1 I_{1j})\end{aligned} \tag{12}$$

Because when the right term of (10) is greater or equal to 0, Theorem 1 is satisfied, we can derive the inequality as

$$\begin{aligned}4H_1 I_{ij} &\geq I_{1i}I_{1j} + H_1 I_{1i} + H_1 I_{1j} \\ &\geq \frac{3}{2} H_1 (I_{1i} + I_{1j})\end{aligned} \tag{13}$$

$I_{1i}I_{1j}$ is always less than or equal to $H_1 I_{1i}$ and $H_1 I_{1j}$ respectively. Thus we can replace $I_{1i}I_{1j}$ with $\frac{1}{2}(H_1 I_{1i} I_{1i} + H_1 I_{1j})$ like second inequality in Eq. (13).

Thus if the right term of (10) is bigger or same than 0, then we can conclude that the error of the Nyström method is bigger than that of the proposed method. This means that when I_{ij} is greater than ${}^{3}/_{4}$ of maximum value between I_{1i} and I_{1j}, (13) is satisfied. Through Theorem 1, we can conclude that when approximating the feature dependency matrix, the error of the proposed method is less than the Nyström method statistically.

We can summarize the proposed feature selection method as follows:

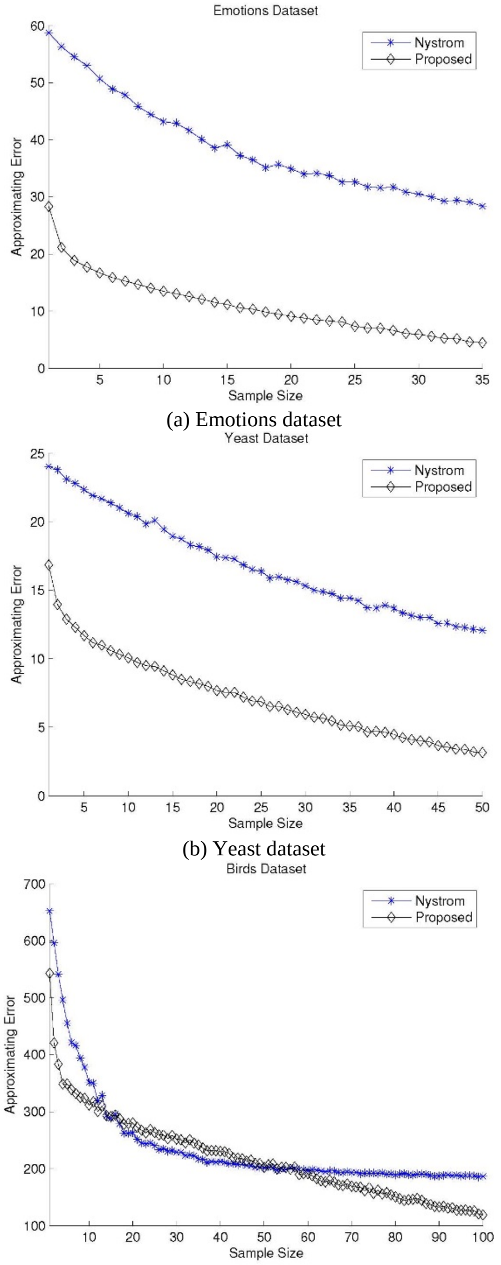
1) Calculate feature dependency $[A\, B]$ of matrix Q and label dependency vector c using MI
2) Approximate E of matrix Q using proposed method
3) Solve the optimization problem of $f(x)$
4) Select the high ranked (weighted) features.

III. Experimental Results

A. Approximation Results

To analyze feature dependency approximation, we compare the proposed method with the Nyström method. In [20], they showed that uniform random sampling technique is the best performance among other sampling techniques for Nyström method. Thus we use uniform random sampling technique, test 100 times for randomness and write the average value. Error value is calculated using Frobenius norm $||\mathrm{Q} - \tilde{Q}||_F^2$. Table I lists the datasets used in our

experiments; they have been widely used for comparative purposes in multi-label classification [21].



(a) Emotions dataset

(b) Yeast dataset

(c) Birds dataset

Fig. 1. Feature dependency approximation error comparison of the proposed method and Nyström method.

Nyström method. Thus we use uniform random sampling technique, test 100 times for randomness and write the average value. Error value is calculated using Frobenius norm $||\mathrm{Q}-\tilde{Q}||_F^2$. Table I lists the datasets used in our experiments; they have been widely used for comparative purposes in multi-label classification [21].

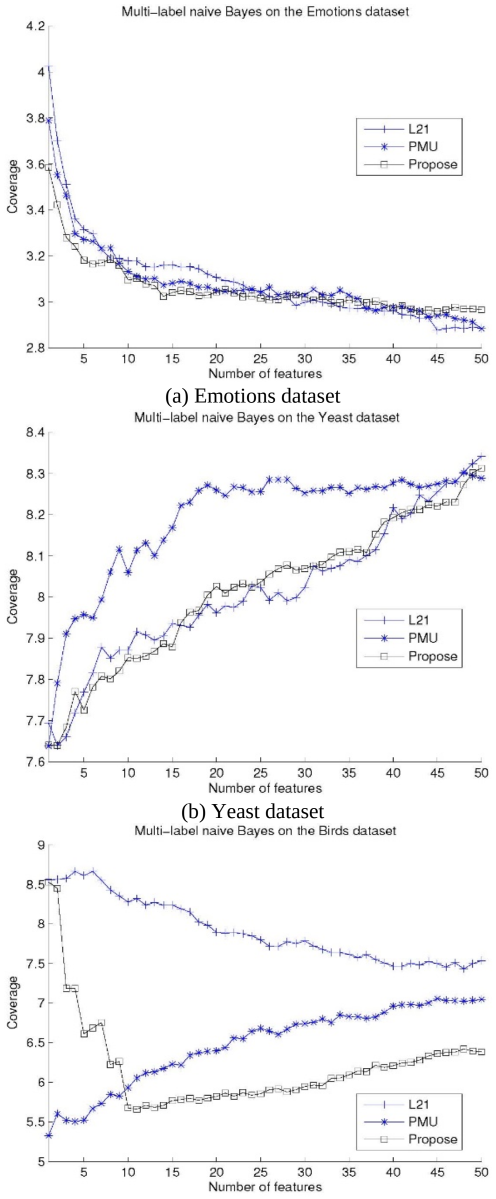
TABLE I: DATA SETS USED IN THE EXPEIRMENTS

Datasets	Patterns	Features	Labels	Domain
Emotions	593	72	6	Music
Yeast	2,417	103	14	Biology
Birds	645	260	19	Audio

Fig. 1 shows the approximation error values of the Nyström method and the proposed method. The vertical axis represents approximation error, and the horizontal axis represents the number of sampled features. As the number of sample increases, the approximating error of two methods decreases. However, reduction ratio of the proposed is bigger than that of the Nyström method. Especially in the Emotions and Yeast datasets, error difference between the proposed method and Nyström method is big. We can conclude that the proposed method is much better than the Nyström method for feature dependency approximation.

TABLE II: EXECUTION TIME COMPARISON

Methods	$L_{2,1}$	PMU	Proposed
Emotions	7.7656	15.9882	**1.2816**
Yeast	56.9317	110.9540	**2.8794**
Birds	**0.6970**	154.2143	5.1645



(a) Emotions dataset

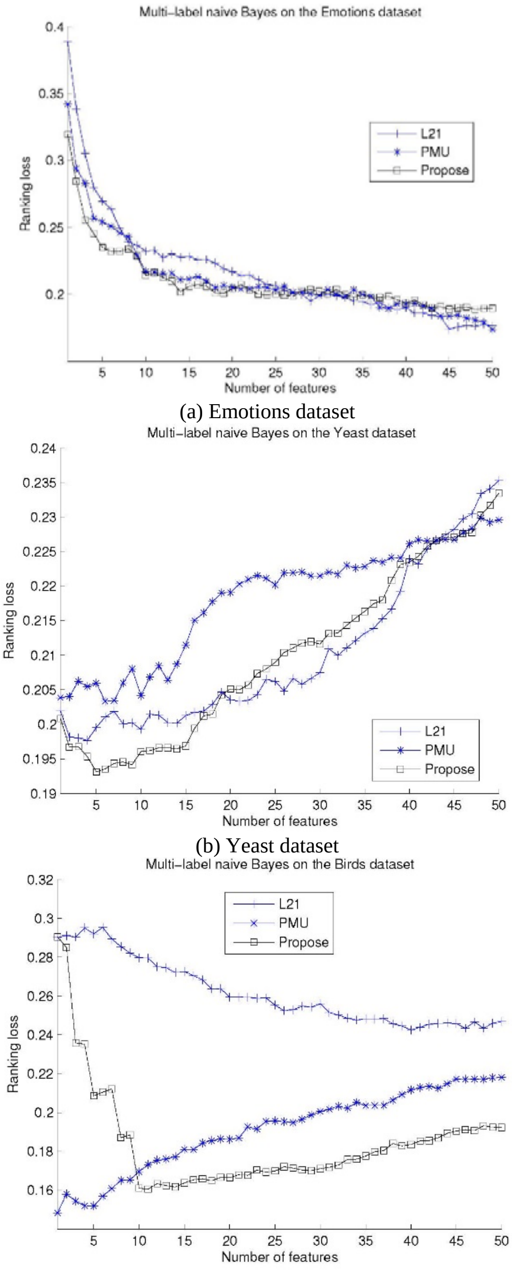
(b) Yeast dataset

(c) Birds dataset

Fig. 2. Coverage comparison of the proposed method conventional methods.

B. Approximating Feature Dependency

We compared the proposed method with conventional multi-label feature selection methods considering the feature dependency [13], [22]. Pairwise multi-label utility (PMU) and $L_{2,1}$. We set the number of iteration for method $L_{2,1}$ to 10. This number is proportional to execution time. Our proposed method needs to sampling ratio. We set the sampling ratio to 0.2 and used uniform random sampling.



(a) Emotions dataset

(b) Yeast dataset

(c) Birds dataset

Fig. 3. Ranking loss comparison of the proposed method conventional methods.

Table II shows the execution time of each feature selection methods. Except for Birds dataset, in two datasets the proposed method outperforms PMU and $L_{2,1}$. In Birds dataset, the reason that $L_{2,1}$ is faster than other method is why the Birds dataset has many 0 values. Calculation cost is low when the sparsity of dataset is high because $L_{2,1}$ needs to matrix inverse calculations in the algorithm. By comparison, the Emotions and Yeast datasets are dense. Compared with PMU, the proposed method is fast about 30 times 12 times, and 39 times in each Birds, Emotions, and Yeast dataset.

Fig. 2 and Fig. 3 show the classification performance, coverage and ranking loss of each method respectively. The low values of coverage and ranking loss represents the high classification performance. We can see that the proposed method outperforms $L_{2,1}$ almost and is similar with PMU. From classification performance comparison, we can conclude that the proposed method shows similar accuracy performance with PMU, but is much faster than that, and speed of the proposed method is similar with $L_{2,1}$, but shows robust accuracy performance than that. We can see that the wide approximation error can cause the difference of feature selection performance.

APPENDIX

Proposition 2: When we know only $[A\ B]$ of matrix Q in Eq. (5) where A ∈ ℝ and B ∈ $\mathbb{R}^{1\times(N-1)}$, the Nyström method approximates E of matrix Q using

$$\mathrm{E} \approx B^T A^{-1} B = \frac{B^T B}{A}$$

In submatrix E of feature dependency matrix Q based on MI, one element Q_{ij} can be approximated by the Nyström method.

Proposition 3: When we know only $[A\ B]$ of matrix Q in Eq. (5) where A ∈ ℝ and B ∈ $\mathbb{R}^{1\times(N-1)}$, the Nyström method approximates one element Q_{ij} of feature dependency matrix Q using

$$Q_{ij} \approx \frac{I_{1i} I_{1j}}{2H_1}$$

Then we can define the error of the Nyström method about one element Q_{ij} for feature dependency approximation.

Lemma 2: When we know only $[A\ B]$ of matrix Q in Eq. (5) where A ∈ ℝ and B ∈ $\mathbb{R}^{1\times(N-1)}$, the approximating error of one element Q_{ij} of the Nyström method for feature dependency can be defined as

$$E_{Nys} = |I_{ij} - \frac{I_{1i} I_{1j}}{2H_1}|$$

REFERENCES

[1] M. Zhang and Z. Zhou, "A review on multi-label learning algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 99, 2013.

[2] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognit.*, vol. 45, no. 9, pp. 3084–3104, 2012.

[3] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," *Lect. Notes Comput. Sci.*, vol. 3201, pp. 217–226, 2004.

[4] R. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2, pp. 135–168, 2000.

[5] M. Boutell, J. Luo, X. Shen, and C. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, 2004.

[6] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music into emotions," in *Proc. 9th Int. Society Music Information Retrieval*, Philadelphia, USA, Sep 2008, pp. 325–330.

[7] Y. Rao, Q. Li, X. Mao, and L. Wenyin, "Sentiment topic models for social emotion mining," *Information Sciences*, vol. 266, pp. 90–100, 2014.

[8] A. N. Srivastava and B. Zane-Ulman, "Discovering recurring anomalies in text reports regarding complex space systems," in *Proc. IEEE Aerospace Conf.*, Big Sky, USA, March 2005, pp. 3853–3862.

[9] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. 7th European Conf. Computer Vision*. Copenhagen, Denmark: Springer, May 2002, pp. 97–112.

[10] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[11] Y. Zhang and Z. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 4, no. 3, p. 14, 2010.

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

[13] J. Lee and D.-W. Kim, "Feature selection for multi-label classification using multivariate mutual information," *Pattern Recognit. Lett.*, vol. 34, 3, pp. 349–357, 2013.

[14] X. Kong and P. Yu, "gMLC: A multi-label feature selection framework for graph classification," *Knowl. Inf. Syst.*, vol. 31 , no. 2, pp. 281–305, 2013.

[15] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang, "Document transformation for multi-label feature selection in text categorization," in *Proc. 7th IEEE Int. Conf. Data Mining*, Omaha, USA, Oct 2007, pp. 451–456.

[16] I. Rodriguez-Lujan, R. Huerta, C. Elkan, and C. S. Cruz, "Quadratic programming feature selection," *The Journal of Machine Learning Research*, vol. 11, pp. 1491–1516, 2010.

[17] A. K. Farahat, A. Ghodsi, and M. S. Kamel, "A novel greedy algorithm for nystr¨om approximation," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2011, pp. 269–277.

[18] Y.-B. Zhao, "An approximation theory of matrix rank minimization and its application to quadratic equations," *Linear Algebra and its Applications*, vol. 437, no. 1, pp. 77–93, 2012.

[19] L. Hyunki, L. Jaesung, and K. Dae-Won, "Multi-label learning using mathematical programming," *IEICE Transactions on Information and Systems*, vol. 98, no. 1, pp. 197–200, 2015.

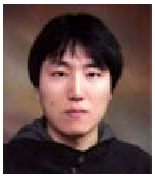
[20] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling techniques for the Nyström method," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2009, pp. 304–311.

[21] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, "Mulan: A java library for multi-label learning," *The Journal of Machine Learning Research*, vol. 12, pp. 2411–2414, 2011.

[22] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint l2;1-norms minimization," *Adv. Neural Inf. Process. Syst.*, vol. 23, pp. 1813–1821, 2010.

[23] Z. Barutcuoglu, R. Schapire, and O. Troyanskaya, "Hierarchical multilabel prediction of gene function," *Bioinformatics*, vol. 22, no. 7, pp. 830–836, 2006.

H. Lim is a Ph.D. candidate in computer science and engineering, Chung-Ang University. He received the M.S. in 2012. He is interested in optimization method and feature selection.



J. Lee received the M.S. and Ph.D. in computer science from Chung-Ang University, Korea in 2009 and 2013, respectively. He participates in post-doctoral course at Chung-Ang University in the School of Computer Science and Engineering, Chung-Ang University in Seoul. His research interest includes biomedical informatics and affective computing. In theoretical domain, he also studies classification, feature selection, and multi-label learning with information theory.

D.-W. Kim received the M.S. and Ph.D. in computer science from KAIST, Korea in 1997 and 2004, respectively. He is currently a professor in the School of Computer Science and Engineering, Chung-Ang University in Seoul, Korea. His research interest includes advanced data mining algorithms with innovative applications to bioinformatics, music emotion recognition, educational data mining, affective computing, and robot interaction.