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 $x \in \mathbb{R}^N$ that contains suitable feature weights; and select the $n$ features with the highest weight values. Because the number of features being selected is limited, 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

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

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

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

$$Q_{ij} = I(f_i, f_j)$$  \hspace{1cm} (2)$$

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

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

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

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

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

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

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 $E$.

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

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 only know $[A B]$ 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

$$Q_{ij} \approx \frac{I_{i1} + I_{1j}}{2}$$  \(7\)

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

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

$$E_{pro} = \left| I_{ij} - \frac{I_{i1} + I_{1j}}{2} \right|$$  \(8\)

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.

$$3(I_{i1} + I_{1j}) \geq I_{ij}$$  \(9\)

The approximating error $E_{Nystr}$ 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.

$$E_{Nystr}^2 - E_{pro}^2 = (E_{Nystr} - E_{pro}) \times (E_{Nystr} + E_{pro})$$  \(10\)

We can derive the left term of (10) as

$$E_{Nystr}^2 - E_{pro}^2 = \left( I_{ij} - \frac{I_{i1} + I_{1j}}{2} \right) \times \left( \frac{I_{i1} + I_{1j}}{2} \right)$$

$$= \frac{1}{2H_1} (H_1I_{i1} + H_1I_{1j} - H_1I_{i1j})$$  \(11\)

The left term of (10) is always greater than or equal to 0 because $H_1 \geq I_{i1}$ and $H_1 \geq I_{1j}$ in information theory. In the same way, we can derive right term of (10) as

$$E_{Nystr} + E_{pro} = 2I_{ij} - \frac{I_{i1} + I_{1j}}{2} - \frac{I_{i1} + I_{1j}}{2}$$

$$= \frac{1}{2H_1} (4H_1I_{i1j} - I_{i1}I_{1j} - H_1I_{i1} - H_1I_{1j})$$  \(12\)

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

$$4H_1I_{i1j} \geq I_{i1}I_{1j} + H_1I_{i1} + H_1I_{1j}$$

$$\geq \frac{3}{2}H_1(I_{i1} + I_{1j})$$  \(13\)

$I_{i1}$, $I_{1j}$ is always less than or equal to $H_1I_{i1}$ and $H_1I_{1j}$ respectively. Thus we can replace $I_{i1}$, $I_{1j}$ with $\frac{3}{2}(H_1I_{i1} + H_1I_{1j})$ 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 $\frac{3}{4}A$ of maximum value between $I_{i1}$ 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 $||Q - \bar{Q}||_F^2$. Table I lists the datasets used in our
experiments; they have been widely used for comparative purposes in multi-label classification [21].

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 $||Q - \tilde{Q}||_F$. Table I lists the datasets used in our experiments; they have been widely used for comparative purposes in multi-label classification [21].

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Patterns</th>
<th>Features</th>
<th>Labels</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotions</td>
<td>593</td>
<td>72</td>
<td>6</td>
<td>Music</td>
</tr>
<tr>
<td>Yeast</td>
<td>2,417</td>
<td>103</td>
<td>14</td>
<td>Biology</td>
</tr>
<tr>
<td>Birds</td>
<td>645</td>
<td>260</td>
<td>19</td>
<td>Audio</td>
</tr>
</tbody>
</table>

Table I: Data Sets Used In The Experiments

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>
<thead>
<tr>
<th>Methods</th>
<th>$L_{2,1}$</th>
<th>PMU</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotions</td>
<td>7.7656</td>
<td>15.9882</td>
<td>1.2816</td>
</tr>
<tr>
<td>Yeast</td>
<td>56.9317</td>
<td>110.9540</td>
<td>2.8794</td>
</tr>
<tr>
<td>Birds</td>
<td>0.6970</td>
<td>154.2143</td>
<td>5.1645</td>
</tr>
</tbody>
</table>

Table II: Execution Time Comparison

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. 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 L2_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 \([AB]\) of matrix \(Q\) in Eq. (5) where \(A \in \mathbb{R}\) and \(B \in \mathbb{R}^{1 \times (N-1)}\), the Nyström method approximates \(E\) of matrix \(Q\) using

\[
E \approx B^T A^{-1} B = \frac{B^T B}{A}
\]

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 \([AB]\) of matrix \(Q\) in Eq. (5) where \(A \in \mathbb{R}\) and \(B \in \mathbb{R}^{1 \times (N-1)}\), the Nyström method approximates one element \(Q_{ij}\) of feature dependency matrix \(Q\) using

\[
Q_{ij} \approx \frac{I_{ij} I_{ij}}{2H_i}
\]

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

\[
E_{NYS} = \left| I_{ij} - \frac{I_{ij} I_{ij}}{2H_i} \right|
\]

### REFERENCES


H. Lim is a Ph.D. candidate in computer science and engineering, Chun-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 Chun-Ang University, Korea in 2009 and 2013, respectively. He participates in post-doctoral course at Chun-Ang University in the School of Computer Science and Engineering, Chun-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, Chun-Ang University in Seoul. His research interest includes advanced data mining algorithms with innovative applications to bioinformatics, music emotion recognition, educational data mining, affective computing, and robot interaction.