Simultaneous analysis of multi-label classification and dimensionality reduction with clustering labels

Hiroaki Ikuta*
Doshisha University, Kyoto, Japan - dio0006@mail4.doshisha.ac.jp

Kensuke Tanioka
Doshisha University, Kyoto, Japan - eim1001@mail4.doshisha.ac.jp

Hiroshi Yadohisa
Doshisha University, Kyoto, Japan - hyadohis@mail.doshisha.ac.jp

Abstract

Recently, multi-label classification has been adopted in domains such as semantic image/text annotation of and functional genomics. However, because multi-label classification tasks involve numerous high-dimensional data, overfitting is a common problem. To solve the overfitting problem, dimensions of the data are reduced before performing the multi-label classification. This two-step method introduces a new problem, because the dimensionality reduction and classification steps are optimized by different criteria. Consequently, the information that is useful for discriminating the multiple labels may be lost after the dimensional reduction, degrading the performance of the discriminant analysis. To resolve this problem, we propose a method that simultaneously achives multi-label classification and dimensionality reduction, but which avoids the pitfals of a previous simultaneous method when there exists several pairs of interrelated label and variable subsets. Our purpose is to detect the label-variable pairs, the low dimensions for the variables, and the variable coefficients that maximize the performance. To this end, we apply different dimensional reduction on each group of prtitiones labels. The superior performance of our proposed method (relative to the existing method) is demonstrated in simulations.

Keywords: alternating least squares; ridge regression; binary relevance.

1. Introduction

Recently, multi-label classification has been adopted in semantic annotation of images and text, functional genomics, and similar domains. In multi-label classification analysis, "each objects is initially allocated to multiple labels, forming the so-called *multi-label data*", is assumed to be given, in which each object is allocated to multiple labels. Next, the labels allocated to each object in these given data are estimated. Read et al. (2011) note that this method is expected to improve the discriminant performance by inter-relating the labels. However, because large multivariate datasets are high-dimensional, they are prone to overfitting as reported by Hawkins (2004). Overfitting may be beneficial for training data, but detrimental to other data, because the number of dimensions is too high for the sample size. The overfitting problem in high-dimensional data is commonly solved by the dimensionality reduction. Methods for dimensionality reduction involve unsupervised learning (such as principal component analysis) and supervised learning (such as canonical correlation analysis). Zhang & Zhou (2010) note that the overfitting problem exsist on multi-label classification. Ji & Ye (2009) report that the overfitting problem has previously been solved by performing discriminant analysis after the dimensionality reduction. However, dimensionality reduction and discriminant analysis are optimized by different criteria. Consequently, the information that is useful for discriminating the multiple labels may be lost after dimensional reduction, thus degrading the performance of the discriminant analysis. To solve this problem, Ji & Ye (2009) simultaneously performed multi-label classification and dimensionality reduction. However, approach of Ji & Ye (2009) aggravates the problem under certain situations; especially, when there exist several pairs of interrelated label and variable subsets. To overcome this problem, we detect each pair, identify the low dimensions for the variables, and find the

variable coefficients that ensure strong performance. To this end, we extend approach of Ji & Ye (2009) by applying different dimensionality reduction to each group of partitioned labels.

2. Related works

This section, discusses previous attempts to simultaneously performe multi-label classification and dimensionality reduction.

Let $\boldsymbol{x}_i \in \mathbb{R}^d$ be the feature vector of the i-th object, $\mathcal{L}$ ($|\mathcal{L}| = k$) be the complete set of labels in the task, and $Y_i \subseteq \mathcal{L}$ be the label set of the i-th object, approach of Ji & Ye (2009) was to minimize the following objective function L_{DC}.

$$L_{DC}(\{f_\ell\},\ \boldsymbol{Q}) = \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} L(f_\ell(\boldsymbol{x}_i),\ y_{i\ell}) + \mu \|\boldsymbol{w}_{\boldsymbol{\ell}}\|_2^2 \right), \tag{1}$$

where $f_\ell(\boldsymbol{x}) = \boldsymbol{w}_\ell' \boldsymbol{Q}' \boldsymbol{x} + b_\ell$ ($\ell = 1,\ 2,\ \ldots,\ k$), $\boldsymbol{w}_\ell \in \mathbb{R}^d$ is a vector of coefficient, $\boldsymbol{Q} \in \mathbb{R}^{d \times r}$ is an orthogonal matrix, $r \leq d$ is the number of reduced dimensions, $b_\ell \in \mathbb{R}$ is a bias term, $\mu \geq 0$ is a regularization parameter, $L(\cdot)$ is a loss function, and

$$y_{i\ell} = \begin{cases} 1 & \ell \in Y_i \\ -1 & \text{otherwise.} \end{cases}$$

Ji & Ye (2009) proposed three forms of the loss function L of Eq. (1); the least square loss function, the hinge loss function and squared hinge loss function. Adopting the least square loss, Eq. (1) becomes

$$L_{\mathrm{LS}}(\{f_\ell\},\ \boldsymbol{Q}) = \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} (f_\ell(\boldsymbol{x}_i) - y_{i\ell})^2 + \mu \|\boldsymbol{w}_\ell\|_2^2 \right). \tag{2}$$

Alternatively, if we assume that both of $\boldsymbol{X} = [\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \ldots,\ \boldsymbol{x}_n]' = (x_{ij}) \in \mathbb{R}^{n \times d}$ and $\boldsymbol{Y} = (y_{i\ell}) \in \mathbb{R}^{n \times k}$ are centered, and specify the bias term $b_\ell = 0$ ($l = 1,\ 2,\ \ldots,\ k$), the optimization problem given by Eq. (2) becomes

$$\min_{\boldsymbol{W}, \boldsymbol{Q}} \|\boldsymbol{X}\boldsymbol{Q}\boldsymbol{W} - \boldsymbol{Y}\|_F^2 + \mu \|\boldsymbol{W}\|_F^2, \tag{3}$$

subject to

$$\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}\ ,$$

where $\|\cdot\|_F$ is the Frobenius norm, $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{W} = [\boldsymbol{w}_1,\ \boldsymbol{w}_2,\ \ldots,\ \boldsymbol{w}_k]$.

The $\boldsymbol{W}$ that satisfies Eq. (3) is given by

$$\boldsymbol{W} = (\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{Q} + \mu\boldsymbol{I})^{-1}\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{Y}\ ,$$

and $\boldsymbol{Q}$ is optimized by solving the following objective function.

$$\max_{\boldsymbol{Q}} \mathrm{tr}((\boldsymbol{Q}'(\boldsymbol{X}'\boldsymbol{X} + \mu\boldsymbol{I})\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{Q}).$$

Because the optimal solution of Eq. (3) is closed-form, Ji & Ye (2009) note that simultaneous multi-label classification and dimensionality reduction should not improve the classification performance.

3. Proposed method

Again, we assume that several pairs of label and variable subsets are interrelated. This section, presents the objective function that ensures good estimation under this condition. The method of Ji & Ye (2009), cannot easily obtain a sigle orthogonal matrix $\boldsymbol{Q}$ that appropriately reduces the dimensions of all labels. To overcome this difficulty, the proposed method extracts the partitions of a label set $\mathcal{L}$, $T_t (t = 1,\ 2,\ \ldots,\ g)$, the orthogonal

matrix $\boldsymbol{Q}_t$ for each T_t, and the coefficient matrix $\boldsymbol{W}$. Assuming a bias term $b_\ell = 0$ ($l = 1,\ 2,\ \ldots,\ k$), the objective function is determined as follows:

$$L_{\mathrm{LSU}}(\{\boldsymbol{w}_\ell\}, \{\boldsymbol{Q}_t\}, \{\boldsymbol{U}\}) = \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} \sum_{t=1}^{g} \left(u_{\ell t} \left(\boldsymbol{w}_\ell' \boldsymbol{Q}_t' \boldsymbol{x}_i - y_{i\ell} \right) \right)^2 + \mu \|\boldsymbol{w}_\ell\|_2^2 \right), \tag{4}$$

subject to the constraints

$$\boldsymbol{Q}_t' \boldsymbol{Q}_t = \boldsymbol{I}\ (t = 1,\ 2,\ \ldots,\ g),\ \boldsymbol{U} = (u_{\ell t}) \in \{0,\ 1\}^{k \times g},\ u_{\ell t} = \begin{cases} 1 & \ell \in T_t \\ 0 & \text{otherwise} \end{cases},\ \sum_{t=1}^{g} u_{\ell t} = 1\ (\ell = 1,\ 2,\ \ldots,\ k).$$

Moreover, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered. Therefore, we can rewrite the optimization problem of Eq. (4) in matrix notation as follows:

$$\min_{\boldsymbol{W}, \{\boldsymbol{Q}_t\}, \boldsymbol{U}} \sum_{t=1}^{g} \|(\boldsymbol{X}\boldsymbol{Q}\boldsymbol{W} - \boldsymbol{Y})\boldsymbol{\Gamma}_t\|_F^2 + \mu \|\boldsymbol{W}\|_F^2, \tag{5}$$

where $\boldsymbol{\Gamma}_t = \mathrm{diag}(u_{1t},\ u_{2t},\ \ldots,\ u_{gt})$. Unlike the optimal solution of Eq. (3), optimal solution of Eq. (5) is not closed-form. Therefore, the proposed method is expected to improve the classification performance by simultaneously solving the multi-label classification and dimensionality reduction. Note that when $g = 1$, Eq. (5) reduces to Eq. (2); in other words, our method is an extension of the method of Ji & Ye (2009). Moreover, the label set should be well separated for a given variable group.

4. Algorithm

To optimize Eq. (5), we apply an alternating least squares method. Given training data $\boldsymbol{X}$ and $\boldsymbol{Y}$, we minimize the objective function (3) by succeesively updating $\boldsymbol{Q}_t$, $\boldsymbol{W}$ and $\boldsymbol{T}$ in each iteration. The proposed algorithm is summarized below:

Step1: Initialize $\boldsymbol{Q}_t$, $\boldsymbol{W}$, and $\boldsymbol{U}$ to random numbers.

Step2: Update $\boldsymbol{Q}_t$ ($t = 1,\ 2, \ldots, g$) by Eq. (7).

Step3: Update $\boldsymbol{W}$ by Eq. (8).

Step4: Update $\boldsymbol{U}$ by Eq. (10).

Step5: Repeat Steps2-4 until Eq. (4) converges.

We now present the updated formulas.

Update $\boldsymbol{Q}_t$

We first, update $\boldsymbol{Q}_t$ by updating the following objective function with $\boldsymbol{W}$ and $\boldsymbol{U}$ fixed:

$$\min_{\{\boldsymbol{Q}_t\}} \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} \sum_{t=1}^{g} \left(u_{\ell t} \left(\boldsymbol{w}_\ell' \boldsymbol{Q}_t' \boldsymbol{x}_i - y_{i\ell} \right) \right)^2 + \mu \|\boldsymbol{w}_\ell\|_2^2 \right). \tag{6}$$

To satisfy $\boldsymbol{Q}_t' \boldsymbol{Q}_t = \boldsymbol{I}$ in Eq. (6), we use the exterior penalty function method, by which Eq. (6) modified as follows:

$$\min_{\{\boldsymbol{Q}_t\}} \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} \sum_{t=1}^{g} \left(u_{\ell t} \left(\boldsymbol{w}_\ell' \boldsymbol{Q}_t' \boldsymbol{x}_i - y_{i\ell} \right) \right)^2 + \mu \|\boldsymbol{w}_\ell\|_2^2 \right) + \lambda \sum_{t=1}^{g} \|\boldsymbol{Q}_t' \boldsymbol{Q}_t - \boldsymbol{I}\|_F^2\ , \tag{7}$$

where $\lambda > 0$ is the penalty coefficient. $\boldsymbol{Q}_t$ is updated by numerically solving Eq. (7).

Update $\boldsymbol{W}$

Next, we update $\boldsymbol{W}$ by solving the following objective function with fixed $\boldsymbol{Q}_t$ and $\boldsymbol{U}$:

$$\min_{\{\boldsymbol{w_\ell}\}} \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} \sum_{t=1}^{g} \left(u_{\ell t} \left(\boldsymbol{w}_\ell' \boldsymbol{Q}_t' \boldsymbol{x}_i - y_{i\ell} \right) \right)^2 + \mu \|\boldsymbol{w}_\ell\|_2^2 \right). \tag{8}$$

Eq. (8) is numerically evaluated to obtain the new $\boldsymbol{W}$.

Update $\boldsymbol{U}$

Finally, we update $\boldsymbol{U}$ by solving the following objective function with fixed $\boldsymbol{Q}_t$ and $\boldsymbol{W}$:

$$\min_{\{u_{\ell t}\}} \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} \sum_{t=1}^{g} \left(u_{\ell t} \left(\boldsymbol{w}_\ell' \boldsymbol{Q}_t' \boldsymbol{x}_i - y_{i\ell} \right) \right)^2 + \mu \|\boldsymbol{w}_\ell\|_2^2 \right), \tag{9}$$

subject to $u_{\ell t} \in \{0,\ 1\}$ and $\sum_{t=1}^{g} u_{\ell t} = 1$ $(\ell = 1,\ 2, \ldots, k)$. At update $\boldsymbol{U}$, objective function is the following.

$$u_{\ell t} = \begin{cases} 1 & t = \arg\min_t \sum_{i=1}^{n} \left(u_{\ell t} \left(\boldsymbol{w}_\ell' \boldsymbol{Q}_t' \boldsymbol{x}_i - y_{i\ell} \right) \right)^2 \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

5. Numerical example

To evaluate the discriminantory performance of our proposed method, we compare the simulation results of our method with those of method of Ji & Ye (2009). The simulation involves three steps. First, the paired training datasets for the pair of ($\boldsymbol{X}_{\text{train}}$ and $\boldsymbol{Y}_{\text{train}}$) and test datasets ($\boldsymbol{X}_{\text{test}}$ and $\boldsymbol{Y}_{\text{test}}$) are generated. Second, the parameters for $\boldsymbol{X}_{\text{train}}$ and $\boldsymbol{Y}_{\text{train}}$, are learned in each method for group numbers $g = 1,\ 2,\ 3$ and 5, and the number of reduced dimensions $r = 3$. Third, the results based on the estimated parameters are evaluated on the test datasets $\boldsymbol{X}_{\text{test}}$ and $\boldsymbol{Y}_{\text{test}}$ by the Hamming loss measure.

Generating the traning data

The training data are defined as $\boldsymbol{X}_{\text{train}} \in \mathbb{R}^{n \times (c+noise)}$ and $\boldsymbol{Y}_{\text{train}} \in \{-1,\ 1\}^{n \times k}$, where $n = 30,\ 60,\ 90$ and 300. The number of variable-correlated label conditions is $c = 6,\ 12$ and 18; the number of noise variables $noise = 1,\ 5,\ 10$ and 50; and the number of labels $k = 10$. These combinations of n, c and $noise$ yield 48 patterns of $\boldsymbol{X}_{\text{train}} \in \mathbb{R}^{n \times (c+noise)}$ and $\boldsymbol{Y}_{\text{train}} \in \{-1,\ 1\}^{n \times k}$. In the training dataset $\boldsymbol{X}_{\text{train}} = [\boldsymbol{v}_1,\ \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{c+noise}] = (x_{ij})$, the $\boldsymbol{v}_j$ $(j = 1,\ 2, \ldots,\ c/2)$, take values from 1 to $n/2$ generated from $N(100,\ 20^2)$ and values from $(n/2) + 1$ to n generated from $N(200,\ 20^2)$. Alternatively $\boldsymbol{v}_j$ $(j = (c/2) + 1,\ (c/2) + 2, \ldots,\ c)$, takes values from 1 to $n/3$ and values from $(2n/3) + 1$ to n generated from $N(200,\ 20^2)$ and values from $(n/3) + 1$ to $2n/3$ generated from $N(100,\ 20^2)$. Values $\boldsymbol{v}_j$ $(j = c + 1,\ c + 2, \ldots,\ c + noise)$ are generated from U(0, 300). In the training dataset, $\boldsymbol{Y}_{\text{train}} = [\boldsymbol{h}_1,\ \boldsymbol{h}_2, \ldots, \boldsymbol{h}_k] = (y_{i\ell})$, $\boldsymbol{h}_\ell$ $(\ell = 1,\ 2, \ldots,\ k/2)$ takes values from 1 to $n/2$ generated from $B(1,\ 0.9)$ and values from $(n/2) + 1$ to n generated from $B(1,\ 0.1)$. Alternatively $\boldsymbol{h}_\ell$ $(\ell = (k/2) + 1,\ (k/2) + 2, \ldots,\ k)$ takes values from 1 to $n/3$ and from $(2n/3) + 1$ to n generated from $B(1,\ 0.1)$ and values from $(n/3) + 1$ to $2n/3$ generated from $B(1,\ 0.9)$. Finally, we genarate $\boldsymbol{h}_\ell$ $(\ell = 1,\ 2, \ldots,\ k)$ by replacing 0 with -1. Figure 1 presents a simple overview of $\boldsymbol{X}_{\text{train}}$ and $\boldsymbol{Y}_{\text{train}}$.

Test data

The test data $\boldsymbol{X}_{\text{test}} \in \mathbb{R}^{n \times (c+noise)}$ and $\boldsymbol{Y}_{\text{test}} \in \{-1,\ 1\}^{n \times k}$ are generated identically to $\boldsymbol{X}_{\text{train}}$ and $\boldsymbol{Y}_{\text{train}}$.

Evaluation of the proposed method

Each case was simulated one hundred times and the average and standard deviation of the results were

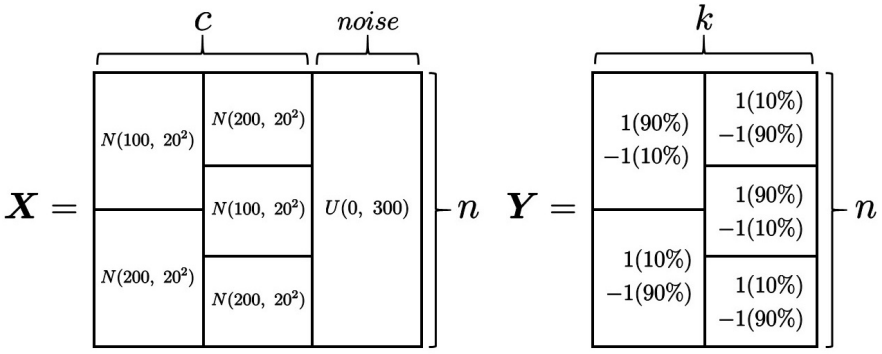

Figure 1: Simple overview of $\boldsymbol{X}_{\text{train}}$ and $\boldsymbol{Y}_{\text{train}}$, where n is the object size, c is the number of variable-correlated label conditions and k is the number of labels

calculated. The results for $n = 30,\ 60,\ 90$ and 300 are summarized in Tables 1, 2, 3 and 4 respectively ($c = 12$ is omitted because of spatial problems). In the proposed method, the average Hamming loss is low for all cases, indicating strong performance. However, increasing the number of groups does not improve the performance of the proposed algorithm.

Table 1: Average and standard diviation of the Hamming loss for $n = 30$

c	*noise*	number of groups			
		1 (previous)	2 (proposed)	3 (proposed)	5 (proposed)
6	1	0.419 ± 0.056	0.345 ± 0.031	0.346 ± 0.048	0.343 ± 0.037
	5	0.377 ± 0.077	0.303 ± 0.038	0.293 ± 0.034	0.290 ± 0.026
	10	0.366 ± 0.064	0.280 ± 0.029	0.275 ± 0.039	0.269 ± 0.025
	50	0.366 ± 0.058	0.255 ± 0.031	0.240 ± 0.026	0.254 ± 0.022
18	1	0.454 ± 0.053	0.289 ± 0.038	0.276 ± 0.030	0.273 ± 0.026
	5	0.418 ± 0.065	0.271 ± 0.026	0.264 ± 0.032	0.263 ± 0.023
	10	0.437 ± 0.070	0.260 ± 0.033	0.249 ± 0.038	0.259 ± 0.020
	50	0.329 ± 0.045	0.252 ± 0.017	0.244 ± 0.039	0.254 ± 0.016

Table 2: Average and standard diviation of the Hamming loss for $n = 60$

c	*noise*	number of groups			
		1 (previous)	2 (proposed)	3 (proposed)	5 (proposed)
6	1	0.405 ± 0.034	0.367 ± 0.027	0.362 ± 0.035	0.364 ± 0.026
	5	0.347 ± 0.026	0.315 ± 0.020	0.310 ± 0.024	0.310 ± 0.018
	10	0.334 ± 0.026	0.293 ± 0.017	0.288 ± 0.022	0.287 ± 0.014
	50	0.420 ± 0.068	0.258 ± 0.014	0.244 ± 0.025	0.255 ± 0.012
18	1	0.420 ± 0.040	0.321 ± 0.025	0.313 ± 0.024	0.313 ± 0.024
	5	0.374 ± 0.041	0.291 ± 0.019	0.280 ± 0.019	0.281 ± 0.018
	10	0.359 ± 0.039	0.280 ± 0.014	0.270 ± 0.020	0.268 ± 0.012
	50	0.387 ± 0.050	0.251 ± 0.014	0.242 ± 0.027	0.254 ± 0.011

Table 3: Average and standard diviation of the Hamming loss for $n = 90$

c	*noise*	number of groups			
		1 (previous)	2 (proposed)	3 (proposed)	5 (proposed)
6	1	0.404 ± 0.028	0.371 ± 0.023	0.364 ± 0.027	0.373 ± 0.022
	5	0.346 ± 0.021	0.320 ± 0.016	0.311 ± 0.016	0.320 ± 0.014
	10	0.328 ± 0.036	0.297 ± 0.015	0.291 ± 0.015	0.294 ± 0.013
	50	0.319 ± 0.021	0.265 ± 0.011	0.251 ± 0.013	0.257 ± 0.017
18	1	0.414 ± 0.031	0.337 ± 0.019	0.326 ± 0.020	0.339 ± 0.023
	5	0.363 ± 0.027	0.301 ± 0.017	0.294 ± 0.016	0.296 ± 0.014
	10	0.336 ± 0.023	0.288 ± 0.014	0.278 ± 0.013	0.279 ± 0.011
	50	0.337 ± 0.025	0.258 ± 0.011	0.245 ± 0.014	0.253 ± 0.010

Table 4: Average and standard diviation of the Hamming loss for $n = 300$

c	*noise*	number of groups			
		1 (previous)	2 (proposed)	3 (proposed)	5 (proposed)
6	1	0.398 ± 0.015	0.381 ± 0.016	0.370 ± 0.018	0.387 ± 0.014
	5	0.340 ± 0.012	0.327 ± 0.012	0.324 ± 0.015	0.331 ± 0.010
	10	0.316 ± 0.008	0.306 ± 0.009	0.302 ± 0.012	0.307 ± 0.007
	50	0.301 ± 0.007	0.284 ± 0.009	0.273 ± 0.012	0.277 ± 0.006
18	1	0.404 ± 0.016	0.369 ± 0.016	0.359 ± 0.014	0.372 ± 0.011
	5	0.342 ± 0.011	0.321 ± 0.012	0.312 ± 0.014	0.323 ± 0.009
	10	0.317 ± 0.009	0.301 ± 0.012	0.293 ± 0.013	0.301 ± 0.007
	50	0.300 ± 0.006	0.281 ± 0.010	0.268 ± 0.012	0.273 ± 0.006

6. Conclusions

The proposed method was expected to improve the discriminatory performance when there exist pairs of interrelated label and variable subsets. The simulation results were not appreciably different among $g = 2$, 3 and 5. Although the proposed method performs better than the method of Ji & Ye (2009), the optimal number of groups appears to be 2. In a future study, the optimal g will be accurately determined. In this paper, the least square loss function is adopted for the proposed objective function. Other loss functions, such as the hinge loss adopted by Ji & Ye (2009), should also be trialed in the proposed method. In addition, we will alter the data structure and apply the method to actual datasets.

References

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, **44**(1), 1–12.

Ji, S., & Ye, J. (2009). Linear Dimensionality Reduction for Multi-label Classification. *In IJCAI*, **9**, 1077–1082.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, **85**(3), 333–359.

Zhang, Y., & Zhou, Z. H. (2010). Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **4**, 14.