PPML: Penalized Partial Least Squares Discriminant Analysis for Multi-Label Learning

Zongjie Ma[1], Huawen Liu[1,2,⋆], Kaile Su[1], and Zhonglong Zheng[1]

[1] Department of Computer Science, Zhejiang Normal University, China
hwliu@zjnu.edu.cn
[2] NCMIS, Academy of Mathematics and Systems Science, CAS, China

Abstract. Multi-label learning has attracted widespread attention in machine learning, and many multi-label learning algorithms have been witnessed. However, two main challenging issues remain: the high dimension of data and the label correlation. In this paper, a new classification method, called penalized partial least squares discriminant analysis for multi-label learning (PPML), is proposed. It aims at performing dimension reduction and capturing the label correlations simultaneously. Specifically, PPML first identifies a latent space for the variable and label space via partial least squares discriminant analysis (PLS-DA). To tackle with the problem of high dimensionality in solving PLS-DA, a ridge penalization is exerted on the optimization problem. After that, the latent space is used to construct learning model. The experimental results on the standard public data sets indicate that PPML has better performance than the state-of-the-art approaches.

Keywords: multi-label learning, partial least squares discriminant analysis, ridge regression, dimension reduction

1 Introduction

Multi-label classification has attracted widespread attention in machine learning, and has been applied to many fields, such as gene function [1], semantic annotation of images [2], and so on [3]. Currently, many learning approaches for multi-label data have been developed. They can be roughly divided into two groups: problem transformation and algorithm adaptation [4]. The former transforms the multi-label problems into a set of single-label problems, which can be solved with the traditional classification algorithms. Binary Relevance (BR) and Label Powerset (LP) [4] are two typical problem transformation approaches. As BR and LP can not work well if the label set has large numbers, Random k-labelsets (RAkEL) [5] employs LP to learn a corresponding classifier by breaking the original set of labels into a number of small random subsets. Nevertheless, RAkEL shows low efficiency when the label set is large and sparse. Algorithm adaptation exploits the traditional learning algorithms to handle the

⋆ Corresponding author.

F. Li et al. (Eds.): WAIM 2014, LNCS 8485, pp. 645–656, 2014.

multi-label problems. For instance, Multi-Label k-Nearest Neighbor (ML-kNN) [6] is based on the k-Nearest Neighbor (kNN) algorithm, which is a lazy learning approach and ignores the label correlations. Backpropagation for Multi-label Learning (BP-MLL) [7] is a neural network approach based on the popular backpropagation algorithm. However, with the increase of the sample dimensions, the efficiency of the learning algorithms becomes a serious problem.

The so-called "curse of dimensionality" resulted from the high dimensionality of data brings enormous challenges to multi-label classification [8]. The existing dimensionality reduction methods are not very appropriate for the multi-label classification problems. Principal component analysis (PCA) [9], which is a classical dimensionality reduction approach and widely used in practice, ignores the label information. Linear discriminant analysis (LDA) [10], a popular supervised dimensionality reduction algorithm, does not consider the dependence between the variables and labels. Canonical correlation analysis (CCA) [11] maximizes the correlation between two blocks of variables, but produces a generalized eigenvalue problem with higher computational expense.

The challenge has been attempted to solve in multi-label classification by researchers during the past few years. For example, Huang and Zhou [12] encode the local influences of label correlations in a LOC code in order to exploit label correlations locally. In [13], a new criterion named PRO LOSS is proposed, which concerns the prediction on all labels and the rankings of only relevant labels. In order to high-dimensional data, Liang Sun et al. [14] structure an equivalent least-squares formulation for CCA under a mild condition. A semi-supervised framework is obtained in [15], which performs optimization for dimension reduction and multi-label inference. Bayesian network structure is used to encode the conditional dependencies of both the labels and feature sets in [16]; Nevertheless, this high-order approach may lead to high model complexities.

In this work, we present a new general framework for multi-label classification, capturing the label correlations and dimensionality reduction simultaneously. The proposed framework is named as penalized partial least squares discriminant analysis for multi-label classification (PPML). It aims at reducing the dimension of muliti-label data and obtaining the latent variables between the variable and label spaces. Specifically, we adopt the technique of partial least squares discriminant analysis (PLS-DA) [17] to discover the latent variables between the variable and label spaces of data, so as to model the correlation between them. To tackle with the problem of "large p, small n", a ridge penalization [18] is further performed on the object optimization function of PLS-DA. After obtaining the latent space, we build a discriminant model and use to predict the label sets for new instances in terms of the regression coefficient of the latent variables.

The rest of the paper is divided into the following sections: Section 2 briefly reviews the state-of-the-arts of multi-label learning. Our proposed method, PPML, is presented in Section 3. Section 4 is the part of experimental comparative study. Finally, we conclude the paper.

2 Related Work

In this section, we briefly review the representative multi-label learning methods. More details can be found in good papers (e.g., [4]).

Generally, multi-label learning algorithms can be divided into two categories: problem transformation and algorithm adaptation. The problem transformation methods firstly transform the multi-label problems into a set of corresponding single label ones, which can be solved by the traditional classification approaches. The characteristic of this kind of methods is fitting the multi-label data to learning algorithms. Binary relevance (BR) is a typical example of this kind methods [4]. It builds a binary classifier for each label occurring in data, where the instances are considered as positive if they contain the label and negative otherwise. Note that BR does not take the correlation of labels into consideration.

CLR [19] and MLStacking [20] exploit the pairwise correlation between the labels to construct classification models. Specifically, they first take a pair of labels as a new label at each time. The instances involve only one of labels in the pair are considered as positive or negative, depending on the containing label. This kind of learning algorithms belong to the second-order transformation methods [21]. Contrastively, RAkEL [5] is a high-order transformation method, where the correlation between multiple labels is involved. However, they have relatively high complexity, and can not handle the high-dimensional data effectively. Other problem transform algorithms, such as Pruned Problem Transformation (PPT) [22] and Classifier Chains (CC) [23] have similar situations.

The second learning methods, i.e., algorithm adaptation, cope with the multi-label learning problems by extending the traditional learning algorithms directly, so as to adapt to the multi-label data. In other words, this kind of methods is fitting learning algorithms to data. As a representative example, ML-kNN extends the traditional kNN learning algorithm, so that it can handle the multi-label data appropriately. ML-DT [24] is another first-order adapting method, where decision tree has been revised according to the properties of multi-label data. Ranking Support Vector Machine (Rank-SVM) [25] is a second-order method, which tries to find maximum margins within the multi-label data. A high-order method named LEAD [16] adopts Bayes learning to deal with the multi-label data. It encodes the conditional dependencies of the variables and labels simultaneously.

How to exploit the correlation of labels and variables of multi-label data is still an open issue for multi-label learning. Although there are some multi-label learning algorithms, exploiting the label dependencies for the multi-label data, they have relatively high complexity, resulting in low robust to high dimensional data. This paper presents a novel approach named PPML. It takes both the correlation of variables and labels and dimension reduction into account simultaneously.

3 PPML Methodology

In this section, we first briefly give the formal concepts of multi-label learning and partial least squares discriminant analysis, and then propose a new learning framework for the multi-label data.

3.1 Multi-Label Classification

Without loss of generality, let $X \in R^{n\times m}$ and $Y \in R^{n\times d}$ be the variable and the label spaces, respectively, where n is the number of instances, m is dimensions of instances and d is the number of class labels involving in instances. In multi-label learning, each instance $x_i \in X(i = 1, 2, ..., n)$ is a vector of variables, and it corresponds to a possible multi-label set $y_i \in Y$, where y_i equals to 1 if the corresponding instance x is tagged with the i-th class label, otherwise $y_i = 0$. Given a data set $D = \{(x_i, y_i) \mid 1 \leq i \leq n\}$ consisting of n multi-label instance, the purpose of multi-label learning is to build a classifier g: $X \rightarrow 2^Y$ from D, and then use this model to predict the labels of unseen instances. From this definition, one may observe that the output of multi-label classification model is a subset of labels, not a single label. This is the distinguished difference of multi-label learning to the traditional ones.

3.2 Partial Least Squares

Partial least squares regression (PLS regression) is a statistical method that finds a linear regression model by projecting the predicted variables and the independent variables to a new space. It shows some similar properties to principal components regression, which tries to find hyper-planes of minimum variance between the predicted and independent variables. PLS regression is particularly suited to the case of high dimensionality, where there is multi-collinearity among the variables. In this case, standard regression will always fail.

PLS tries to locate the fundamental relations between two matrices, i.e. a latent variable approach to modeling the covariance structures in these two spaces. Assume that $X \in R^{n\times m}$ and $Y \in R^{n\times d}$ denote the independent and predicted variables. Usually, they can be decomposed by the common latent components $T \in R^{n\times k}$ as follows:

$$X = TP^T + E \tag{1}$$

$$Y = TQ^T + F \tag{2}$$

where $T = (t_1, t_2, ..., t_k) \in R^{n\times k}$ denote the score (latent) vectors of X and Y, i.e., the latent space. $P = (p_1, p_2, ..., p_k) \in R^{m\times k}$ and $Q = (q_1, q_2, ..., q_k) \in R^{d\times k}$ are the loading vectors of X and Y, respectively. $E \in R^{n\times m}$ and $F \in R^{n\times d}$ are the residual matrixes.

According to the linear transformation, we make an assumption that the latent components T is a linear tramsformation of X as follows:

$$T = XW \tag{3}$$

where $W \in R^{m \times k}$ is the weight matrix. After T is constructed, Q^T in Eq.(2) can be obtained by solving the least squares problem, i.e.,

$$Q^T = (T^T T)^\dagger T^T Y \tag{4}$$

where $(T^T T)^\dagger$ is the Moore-Penrose inverse of $T^T T$. Substituting (3) into (2), for Y we have its regression form:

$$Y = XB + F \tag{5}$$

where the regression coefficient B is $B = WQ^T = W(T^T T)^\dagger T^T Y$.

From the definitions above, we know that a PLS model tries to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. Depending on the tasks, PLS have many variants. For example, partial least squares discriminant analysis (PLS-DA) is widely used when the Y is categorial.

Given a new instance X_{new}, its categories or labels can be predicted in terms of the PLS model (i.e., Eq.(5)) as follows:

$$\hat{Y} = X_{new} B = X_{new}(WQ^T) \tag{6}$$

PLS-DA is an efficient dimension reduction tool for handling the multi-label data. However it is not a specifical method for the cases of "large p, small n" and feature selection. Besides, the results obtained by PLS-DA are often difficult to be interpreted. Just for this reason, we impose a ridge penalization on the PLS-DA model to alleviate this problem.

3.3 PLS-DA with Ridge Penalization

For the high dimensionality of multi-label data, obtaining the PLS-DA model in a straightforward way becomes unfeasible, especially when the variables are highly collinear.

According to the Eq.(1) and (2), one may observe that PLS aims at locating the common latent variables T of X and Y, such that their covariance is maximal. That is to say, the object function of PLS-DA can be represented in a equivalent form as follows:

$$\begin{aligned} &argmax_{p,q}\ cov(Xp, Yq) \\ &s.t.\quad \|p\| = 1, \|q\| = 1 \end{aligned} \tag{7}$$

where $cov(Xp, Yq)$ is the covariance of Xp and Yq, p and q are the loading vectors of X and Y respectively.

The optimization problem of Eq.(7) can be solved through the following Lagrange function:

$$L(p, q) = p^T X^T Y q - \frac{\beta}{2}(p^T p - 1) - \frac{\theta}{2}(q^T q - 1) \tag{8}$$

where β and θ are Lagrange multipliers. After differentiating Eq.(8) with respect to p and q respectively, and let them equal to zero, we have the following

equivalent problems:

$$X^T Y Y^T X p = \lambda p \tag{9}$$

$$Y^T X X^T Y q = \lambda q \tag{10}$$

Thus, the optimization problem is now transformed into the problem of solving eigenvalues and eigenvectors. Note that the vectors p and q are the eigenvectors of $X^T Y Y^T X$ and $Y^T X X^T Y$, respectively. An intuitive way of obtaining the eigenvectors p and q is to perform the technique of singular value decomposition (SVD) on $X^T Y$:

$$X^T Y = P \Sigma Q^T \tag{11}$$

where the eigenvectors $P \in R^{m \times r}$ and $Q^T \in R^{r \times d}$ are orthogonal, $\Sigma \in R^{r \times r}$ is the diagonal matrix consisting of the singular values. (p, q) is a pair of eigenvectors corresponding to the eigenvalue in Σ.

It should be mentioned that Eq.(11) may be ill-posed as the dimensionality of data is larger than the number of data. Thus, it should be penalized for the consideration of numerical computing and practical applications. Here we exert a l_2-norm penalty on Eq.(11). This is also known as ridge regularization, which is a method for solving badly conditioned linear regression problems. The benefits of ridge regularization are most striking in the presence of multi-collinearity and penalize the size of the regression coefficients, resulting in shrinking the regression coefficients toward zero [26], [27]. Sparse property is more prefer because it can yield easily interpretable results. Moreover, with the increase of the number of labels, the label space Y is usually sparse. Thus, it is necessary to shrinking the loading vectors of Y.

After applying l_2-norm penalty, the loading vector q of Y can be obtained by solving the constraint optimization problem as follows:

$$q^{ridge} = argmin_q \left\| q^T M - p \right\|_2^2 + \lambda \left\| q \right\|_2^2 \tag{12}$$

where $M = Y^T X$, λ is the regularization parameter for the loading vector q. For q, when λ is enough small, the weight coefficients of some variables compressed to zero by comparing with a threshold. Let $L(q)$ be the Lagrange function of Eq.(12), we have

$$\begin{aligned} L(q) &= \left\| q^T M - p \right\|_2^2 + \lambda \left\| q \right\|_2^2 \\ &= q^T M M^T q - 2 q^T M p^T + p p^T + \lambda q^T q \end{aligned} \tag{13}$$

After taking the derivative of Eq.(13) with respect to q and setting it to zero, we can obtain q^{ridge} as

$$q^{ridge} = [M M^T + \lambda I]^{\dagger} M p^T \tag{14}$$

Substituting M in Eq.(14) with Eq.(11), we further have

$$\begin{aligned} q^{ridge} &= P(\Sigma^2 + \lambda I)^{\dagger} \Sigma Q^T p^T \\ &= \sum_{i=1}^{d} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} P_i (Q_i^T p^T) \\ &= \sum_{i=1}^{d} f(\sigma_i) P_i (Q_i^T p^T) \end{aligned} \tag{15}$$

Alg. 1 The framework of PPML for the multi-label data

Input:
X, Y: The training data with the variable and label spaces
$\bar{X}$: The new instances
λ: The regularization parameter
h: The number of iterations
θ: The threshold value for prediction
Output:
$\bar{Y}$: the predicted labels of $\bar{X}$
Training:
Initialize T, P, Q and W as T=[ ] ,P=[ ],Q=[ ],W=[ ]
For i=1 to h
Obtain the initial values of p and q according to Eq.(11)
Repeat
Obtain p according to q, and normalize it
Obtain q according to Eq.(14), and normalize it
$u = Yq$
Compute w as $w = X^T u$, and normalize it
$t = Xw$
Until convergence
Update X,Y as $X = X - tp^T$; $Y = Y - tq^T$
Update T, P, Q and W as:
$T \leftarrow [T, t], P \leftarrow [P, p], Q \leftarrow [Q, q], W \leftarrow [W, w]$
End For
Predicting:
Compute the real-valued outputs space of $\bar{X}$ according to Eq.(6)
$O = \bar{X}WQ^T$
Compute the final predictive labels space $\bar{Y}$:
$\bar{Y}_{ij} = \begin{cases} 1, & O_{ij} \geq \theta \\ 0, & O_{ij} < \theta \end{cases}$

where $f(\sigma_i)$ is the shrinkage factors.

Based on the analysis above, we propose a new multi-label learning framework called PPML (Penalized Partial least squares discriminant analysis for Multi-label Learning). As the name indicates, our method employs PLS-DA with ridge regularization to handle the classification problem of high-dimensional multi-label data. Algorithm 1 presents the framework of PPML in detail. It exploits Nonlinear Iterative Partial Least Squares (NIPALS) [17] to obtain p and q. Alternatively, PPML can also be implemented with other forms like PLS-SB [28] and SIMPLS [29].

PPML works in a straightforward way and can be easily understood. It mainly consists of two stages, i.e., model training and result predicting. Specifically, in the training stage, two loops are nested. The major purpose of the inner loop is to get the loading vectors p and q of the variable and label spaces respectively, while the outer loop aims at yielding all loading vectors (P and Q), the latent components T and the coefficients W, so as to build the PPML learning model with PLS-DA. In the predicting stage, the prediction value of a new instance is a real-value vector in terms of (6). Later the output will be transformed into a vector of $\{0, 1\}$ by comparing with a given threshold θ, which is often empirically set as 0.5.

Table 1. General information of the experimental data sets

Datasets	Inst.	Var.	Labels	L.Card.	L.Dens.
Medical	978	1449	45	1.245	0.028
Arts	5000	462	26	1.636	0.063
Entertainment	5000	640	21	1.420	0.068
Health	5000	612	32	1.662	0.052
Recreation	5000	606	22	1.423	0.065
Reference	5000	793	33	1.169	0.036
Science	5000	743	40	1.451	0.036

4 Experiments

4.1 Data Sets

In our experiments, seven public data sets from the real-world applications were adopted. They are *Medical*, *Arts*, *Entertainment*, *Health*, *Recreation*, *Reference*, and *Science*. The *Medical* data set was used in the Medical Natural Language Processing Challenge[1] in 2007. In this data set, each instance contains brief free-text summary of a patient symptom history. The last six benchmark data sets were collected from Yahoo. They cover different domains in web page categorization.

Table 1 summaries the general information of the benchmark data sets used in experiments, where *Inst.* and *Var.* denote the number of instances and the dimensionality of data for each data set respectively. In addition, *L.Card.*, representing label cardinality, is the average number of labels per instance, while *L.Dens.*, standing for label density, is the fraction of the cardinality according to the number of labels.

4.2 Comparison of Algorithms

To demonstrate the effectiveness of our algorithm, nine multi-label learning algorithms have been adopted in comparing with PPML. They are BP-MLL [7], BRkNN [30], IBLR_ML [31], LP [4], MLkNN [6], PPT [22], CC [23], MLStacking(MLS) [20], and MAHR [33]. They are representatives of the state-of-the-art multi-label learning algorithms, and stand for different learning manners. They can deal with the multi-label problems and have relatively better performance and efficiency.

The performance of the learning algorithms heavily relies on their parameters. In our experiments, default value was assigned for each parameter as did in the MULAN software package[2]. MULAN [32] is an open source Java library for multi-label learning. It brings many popular multi-label learning algorithms together. For the MAHR classifier, its parameters was set as recommended by

[1] `http://www.computationalmedicine.org/challenge/`
[2] `http://mlkd.csd.auth.gr/multilabel.html`

the authors in the literature [33], that is, the number of boosting rounds was two times of variables for each data set.

4.3 Evaluation Metrics

Multi-label classification needs more complex evaluation metrics than traditional classification. In order to roundly evaluate the performance of PPML and other algorithms, we took four commonly used evaluation metrics. They are *Ranking Loss*, *One-Error*, *Coverage* and *Average Precision* [4].

Ranking Loss (RL) indicates the mis-ordered degree of couples of labels, where an irrelevant label has higher rank than a relevant one. *One-Error* (OE) estimates how many times the top-ranked label is irrelevant to the true class labels for each instance. *Coverage* (Cov) obtains the number of the steps that are needed, on average, to move down the ranked list of labels, in order to cover the whole relevant labels of the instance. *Average Precision* (AP) evaluates the average fraction of true labels ranked above a particular label.

Since *Ranking Loss*, *One-Error* and *Coverage* evaluate the loss of the prediction results, the smaller the metric values, the better the performance of learning algorithms. On the contrary, for *Average Precision*, the larger value indicates the better performance.

4.4 Experimental Results

In the experiments, 10-folds cross-validation was performed on each combination of classifier and data set. The experiments were carried out under the platform of MULAN. Table 2 shows the comparison results of classification performance of classifiers in terms of four evaluation metrics, where the mean value of each algorithm was ecorded on each data set.

From the experimental results in Table 2, one can notice that PPML is promising. It has better performance than others in most cases. For example, PPML achieved the best performance on five over seven data sets at the aspect of *Ranking Loss*. Even on the *Health* and *Science* data sets, the performance of PPML is just slightly worse than the corresponding best one, not the worst one.

Similar situations also present on the *One-Error* and *Coverage* metrics, where PPML outperformed other popular multi-label learning algorithms on six and five over seven benchmark data sets respectively. The one-error of PPML on *Medical* is 13.61%, which is slight higher than that of MAHR.

For the measure of *Average Precision*, PPML is the best in comparing with other classifiers. The performance of PPML is predominant and significantly better than the rest learning algorithms over all of the seven benchmark data sets. For example, on the *Arts* and *Recreation* data sets, the average precisions of PPML are 60.59% and 62.14% respectively, while the highest precisions of other classifiers are 50.92% and 52.01%, achieved by MLS and MAHR respectively.

Table 2. Experimental results of classifiers on four evaluation metrics, where ↓ means the smaller, the better, and ↑ means the larger, the better. Bold value shows the winner on each dataset.

	PPML	BPMLL	BRkNN	IBLR_ML	LP	MLkNN	PPT	CC	MLS	MAHR
					Ranking Loss(↓)					
Medical	**0.0175**	0.4321	0.0596	0.0890	0.1277	0.0573	0.1067	0.0990	0.0910	0.0410
Arts	**0.1379**	0.4545	0.2331	0.1602	0.3946	0.1575	0.2951	0.2512	0.1610	0.1886
Entertainment	**0.1165**	0.3763	0.3858	0.1395	0.4838	0.1364	0.3353	0.2364	0.1399	0.1630
Health	0.0660	0.2469	0.1953	0.0604	0.4310	0.0723	0.2155	0.1316	**0.0611**	0.0656
Recreation	**0.1464**	0.5634	0.2475	0.1811	0.4381	0.1804	0.3265	0.2605	0.1797	0.2513
Reference	**0.0826**	0.2694	0.2472	0.0918	0.4545	0.0889	0.2896	0.1675	0.0905	0.1206
Science	0.1159	0.4564	0.2023	0.1145	0.4480	0.1177	0.3729	0.2311	**0.1139**	0.2182
					One-Error(↓)					
Medical	0.1361	0.9674	0.2347	0.2918	0.1572	0.2388	0.1949	0.2123	0.3081	**0.1123**
Arts	**0.4788**	0.9844	0.8410	0.6124	0.7194	0.6338	0.6612	0.6674	0.6142	0.5801
Entertainment	**0.4368**	0.9582	0.7342	0.6044	0.6294	0.6360	0.5360	0.5356	0.6116	0.4542
Health	**0.2664**	0.9936	0.7228	0.4094	0.5288	0.4598	0.4314	0.3816	0.4138	0.3080
Recreation	**0.4726**	0.9778	0.7670	0.6466	0.6648	0.6664	0.6158	0.6232	0.6462	0.5470
Reference	**0.3838**	0.9644	0.8964	0.4908	0.5864	0.4838	0.5206	0.5476	0.4928	0.4160
Science	**0.5238**	0.9885	0.6244	0.5874	0.7244	0.5940	0.7096	0.6568	0.5876	0.5660
					Coverage(↓)					
Medical	**1.3174**	20.3123	3.9756	5.3643	7.9194	3.7633	6.4061	6.1643	5.5163	2.9185
Arts	**5.3223**	13.0170	7.7288	5.6838	12.1808	5.5830	9.6152	8.6368	5.6956	6.6663
Entertainment	**3.2930**	8.3088	9.0084	3.6808	10.9336	3.5900	7.9446	6.0716	3.6912	4.4364
Health	3.7082	9.1972	8.6658	**3.2216**	16.2388	3.6586	9.1798	6.6592	3.2486	3.6703
Recreation	**4.1814**	12.9760	6.2270	4.7388	10.7832	4.7090	8.2528	6.8510	4.7050	6.5649
Reference	**3.3472**	9.1452	8.9134	3.5550	15.6426	3.4244	10.1608	6.1446	3.4960	4.7160
Science	6.1550	19.6508	10.1670	5.9508	20.0474	5.9376	17.0720	11.7990	**5.9304**	10.9840
					Average Precision(↑)					
Medical	**0.8871**	0.1124	0.7958	0.7497	0.7923	0.8039	0.7967	0.8029	0.7421	0.8791
Arts	**0.6059**	0.1344	0.3752	0.5098	0.3554	0.5032	0.4294	0.4517	0.5092	0.5055
Entertainment	**0.6599**	0.1860	0.4044	0.5445	0.3959	0.5341	0.5169	0.5503	0.5408	0.6196
Health	**0.7707**	0.1817	0.3712	0.6896	0.4719	0.6485	0.6200	0.6805	0.6858	0.7504
Recreation	**0.6214**	0.1210	0.3767	0.5016	0.3925	0.4899	0.4731	0.4953	0.5022	0.5201
Reference	**0.7004**	0.1613	0.3080	0.6064	0.4260	0.6110	0.5342	0.5647	0.6080	0.6441
Science	**0.5809**	0.0849	0.4698	0.5305	0.2938	0.5228	0.3556	0.4382	0.5306	0.4844

5 Conclusions

In this paper, a new multi-label learning framework, called PPML, is proposed to deal with the multi-label problems. It mainly exploits partial least squares discriminant analysis (PLS-DA) to achieve the purpose of performing dimension reduction and capturing the label correlations simultaneously. To cope with the multi-collinearity problem resulted from the high dimensionality of data, a ridge regularization penalty is further exerted on the object optimization function of PLS-DA. The experimental results are encouraging and show that PPML is promising in comparison with the other state-of-the-art algorithms.

In the future, we will make an attempt to find other more efficient methods and combine them with PLS-DA to tackle with the problems of multi-label learning.

Acknowledgements. The authors are grateful to the anonymous referees for their valuable comments and suggestions. This work is partially supported by the National NSF of China (61100119, 61170108, 61170109, 61272130, and 61272468), the NSF of Zhejiang province (LY14F020012), Postdoctoral Science Foundation of

China (2013M530072), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (201204214).

References

1. Barutcuoglu, Z., Schapire, R., Troyanskaya, O.: Hierarchical multi-label prediction of gene function. Bioinformatics 22(7), 830–836 (2006)
2. Yang, S., Kim, S., Ro, Y.: Semantic home photo categorization. IEEE Transactions on Circuits and Systems for Video Technology 17, 324–335 (2007)
3. Read, J.: Scalable multi-label classification, PhD thesis, University of Waikato, Hamilton, New Zealand (2010)
4. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, 2nd edn., Spring (2010)
5. Tsoumakas, G., Katakis, I., Vlahavas, I.: Randomk-Labelsets for Multi-Label Classification. IEEE Transactions on Knowledge and Data Engineering 23(7), 1079–1089 (2011)
6. Zhang, M.-L., Zhou, Z.-H.: ML-kNN: a lazy learning approach to multi-label learning. Pattern Recognition 40(7), 2038–2048 (2007)
7. Zhang, M.-L., Zhou, Z.-H.: Multi-label Neural Network with Applications to Functional Genomics and Text Categorization. IEEE Transactions on Knowledge and Data Engineering 18(10), 1338–1351 (2006)
8. Bellman, R.: Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton (1961)
9. Jolliffe, I.: Principal Component Analysis. Springer, New York (1986)
10. Fisher, R.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188 (1936)
11. Hotelling, H.: Relations between two sets of variables. Biometrika 28, 312–377 (1936)
12. Huang, S.-J., Zhou, Z.-H.: Multi-label learning by exploiting label correlations locally. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012), Toronto, Canada, pp. 949–955. AAAI Press (2012)
13. Xu, M., Li, Y.-F., Zhou, Z.-H.: Multi-Label Learning with PRO Loss. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013), Bellevue, WA (2013)
14. Sun, L., Ji, S.-W., Ye, J.-P.: Canonical Correlation Analysis for Multilabel Classification: A Least-Squares Formulation. Extensions, and Analysis, Pattern Analysis and IEEE Transactions on Machine Intelligence 33(1), 194 (2011)
15. Qian, B., Davidson, I.: Semi-supervised dimension reduction for multi-label classification. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence (2010)
16. Zhang, M.-L., Zhang, K.: Multi-label learning by exploiting label dependency. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010), Washington, D.C., pp. 999–1007 (2010)
17. Wold, H.: Path Models with Latent Variables: The NIPALS Approach. In: Blalock, H.M., Aganbegian, A., Borodkin, F.M., Boudon, R., Capecchi, V. (eds.) Quantitative Sociology: International Perspectives on Mathematical and Statistical Modeling, pp. 307–357. Academic Press, New York (1975)

18. Hoerl, A., Kennard, R.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1), 55–67 (1970)
19. Frnkranz, J., Hllermeier, E., Menca, E.L., Brinker, K.: Multilabel classification via calibrated label ranking. Machine Learning 23(2), 133–153 (2008)
20. Tsoumakas, G., Dimou, A., Spyromitros, E., Mezaris, V., Kompatsiaris, I., Vlahavas, I.: Correlation-based pruning of stacked binary relevance models for multi-label learning. In: Proceedings of the Workshop on Learning from Multi-Label Data (MLD 2009), pp. 101–116 (2009)
21. Zhang, M.-L., Zhou, Z.-H.: A review on multi-label learning algorithms. In: IEEE Transactions on Knowledge and Data Engineering (2013) doI:10.1109/TKDE.2013.39
22. Read, J.: A pruned problem transformation method for multi-label classification. In: Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008), pp. 143–150 (2008)
23. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: ECML/PKDD 2009, pp. 254–269 (2009)
24. Clare, A.J., King, R.D.: Knowledge Discovery in Multi-label Phenotype Data. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 42–53. Springer, Heidelberg (2001)
25. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Dietterich, G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems, vol. 14, pp. 681–687. MIT Press, Cambridge (2002)
26. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research 11(1), 19–60 (2010)
27. Jenatton, R., Mairal, J., Obozinski, G., Bach, F.: Proximal methods for sparse hierarchical dictionary learning. In: Proceedings of the International Conference on Machine Learning, ICML (2010)
28. Sampson, P., Streissguth, A., Barr, H., Bookstein, F.: eurobehavioral effects of prenatal alcohol: Part II. Partial Least Squares Analysis, Neurotoxicology and Teratology 11(5), 477–491 (1989)
29. De Jong, S.: SIMPLS: an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18(3), 251–263 (1993)
30. Spyromitros, E., Tsoumakas, G., Vlahavas, I.P.: An Empirical Study of Lazy Multilabel Classification Algorithms. In: Darzentas, J., Vouros, G.A., Vosinakis, S., Arnellos, A. (eds.) SETN 2008. LNCS (LNAI), vol. 5138, pp. 401–406. Springer, Heidelberg (2008)
31. Cheng, W., H*ü*llermeier, E.: Combining instance-based learning and logistic regression for multilabel classification. Machine Learning 76(2-3), 211–225 (2009)
32. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A Java Library for Multi-Label Learning. Journal of Machine Learning Research 12, 2411–2414 (2011)
33. Huang, S.-J., Yu, Y., Zhou, Z.-H.: Multi-label hypothesis reuse. In: Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Beijing, China, pp. 525–533 (2012)