Multi-Label Dimensionality Reduction via Dependence Maximization

Yin Zhang and **Zhi-Hua Zhou** *
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
{zhangyin, zhouzh}@lamda.nju.edu.cn

Abstract

Multi-label learning deals with data associated with multiple labels simultaneously. Like other machine learning and data mining tasks, multi-label learning also suffers from the curse of dimensionality. Although dimensionality reduction has been studied for many years, multi-label dimensionality reduction remains almost untouched. In this paper, we propose a multi-label dimensionality reduction method, MDDM, which attempts to project the original data into a lower-dimensional feature space maximizing the dependence between the original feature description and the associated class labels. Based on the Hilbert-Schmidt Independence Criterion, we derive a closed-form solution which enables the dimensionality reduction process to be efficient. Experiments validate the performance of MDDM.

Introduction

In many real-world problems one object usually inheres multiple concepts simultaneously. One label per instance is out of its capability to describe such scenario, and thus *multi-label learning* has attracted much attention. Under the framework of multi-label learning, each instance is associated with multiple labels indicating the concepts the instance belongs to. Multi-label learning techniques have already got diverse applications (Yu, Yu, & Tresp 2005; Zhang & Zhou 2007).

The *curse of dimensionality* often causes serious problems to learning with high-dimensional data, and thus lots of dimensionality reduction methods have been developed. Upon whether the label information is used or not, current methods can be roughly classified into two categories, i.e., *unsupervised*, e.g. principal component analysis (PCA) (Jolliffe 1986), or *supervised*, e.g. linear discriminant analysis (LDA) (Fisher 1936). In spite of the fact that multi-label learning tasks usually involve high-dimensional data, multi-label dimensionality reduction remains almost untouched. Direct application of existing unsupervised dimensionality reduction methods to multi-label tasks ignores the label information. As for existing single-label supervised dimensionality reduction methods, one possible way to extend to multi-label learning is to treat every combination of concepts as a class. Such an extension, however, suffers from the explosion of the possible combination of labels. To the best of our knowledge, the only relevant work is the MLSI method described in (Yu, Yu, & Tresp 2005), which is a multi-label extension of Latent Semantic Indexing (LSI). It has been shown that MLSI works well on text categorization tasks (Yu, Yu, & Tresp 2005).

In this paper, we propose a multi-label dimensionality reduction method called MDDM (Multi-label Dimensionality reduction via Dependence Maximization) which tries to identify a lower-dimensional feature space maximizing the dependence between the original feature description and class labels associated with the object. We derive a closed-form solution for MDDM, which enables the multi-label dimensionality reduction process to be not only effective but also efficient. The superior performance of the proposed MDDM method is validated in experiments.

The MDDM Method

Let $\mathcal{X} = \mathbb{R}^D$ denote the feature space and Θ denote a concept set. The proper concepts associated with an instance $\boldsymbol{x}$ is a subset of Θ, which can be represented as a $|\Theta|$-dimensional binary label vector $\boldsymbol{y}$, with 1 indicating that the instance belongs to the concept corresponding to the dimension and 0 otherwise. All the possible labels make up the label space $\mathcal{Y} = \{0,1\}^{|\Theta|}$. Given a multi-label data set $S = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$, the goal is to learn from S a function $h : \mathcal{X} \to \mathcal{Y}$ which is able to predict proper labels for unseen instances.

Motivated by the consideration that there should exist some relation between the feature description and labels associated with the same object, we attempt to find a lower-dimensional feature space in which the dependence between the features and labels are maximized. For simplicity, here we consider a linear projection P, while a non-linear extension can be obtained easily by transforming the primal problem into its dual form and then applying the kernel trick. Assume that the instance $\boldsymbol{x}$ is projected into the new space $\mathcal{F}$ by $\phi(\boldsymbol{x}) = P^T\boldsymbol{x}$. Then, we try to maximize the dependence between the feature description $\phi(\boldsymbol{x}) \in \mathcal{F}$ and the class labels $\boldsymbol{y} \in \mathcal{Y}$. Many criteria can be used to measure such dependence and here we adopt the Hilbert-Schmidt In-

*This research was supported by the National High Technology Research and Development Program of China (2007AA01Z169) and National Science Foundation of China (60635030, 60721002).

dependence Criterion (HSIC) (Gretton *et al.* 2005) due to its simplicity and neat theoretical properties.

An empirical estimate of HSIC (Gretton *et al.* 2005) is

$$\text{HSIC}(\mathcal{F}, \mathcal{Y}, \mathbf{P}_{\boldsymbol{xy}}) = (N-1)^{-2}\text{tr}\left(\mathbf{HKHL}\right), \quad (1)$$

where $\mathbf{P}_{\boldsymbol{xy}}$ is the joint distribution and tr$(\cdot)$ is the trace of matrix. $\mathbf{K} = [K_{ij}]_{N\times N}$ and $\mathbf{L} = [L_{ij}]_{N\times N}$ are the matrices of the inner product of instances in $\mathcal{F}$ and $\mathcal{Y}$ which could also be considered as the kernel matrices of $\mathcal{X}$ and $\mathcal{Y}$ with kernel functions $k(\boldsymbol{x}, \boldsymbol{x}') = \langle\phi(\boldsymbol{x}), \phi(\boldsymbol{x}')\rangle = \langle P^T\boldsymbol{x}, P^T\boldsymbol{x}'\rangle$ and $l(\boldsymbol{y}, \boldsymbol{y}') = \langle\boldsymbol{y}, \boldsymbol{y}'\rangle$. $\mathbf{H} = [H_{ij}]_{N\times N}$, $H_{ij} = \delta_{ij} - 1/N$, δ_{ij} takes 1 when $i = j$ and 0 otherwise.

Since the normalization term in Eq. 1 does not affect the optimization procedure, we can drop it and only consider tr$(\mathbf{HKHL})$. Denote $X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ and $Y = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N]$. Thus $\phi(X) = P^TX$ and $\mathbf{K} = \langle\phi(X), \phi(X)\rangle = X^TPP^TX$, $\mathbf{L} = Y^TY$. We can rewrite the optimization as searching for the optimal linear projection

$$P^* = \arg\max_P \text{ tr}\left(\mathbf{H}X^TPP^TX\mathbf{HL}\right). \quad (2)$$

Suppose we will reduce to a d-dimensional space and denote $P = [P_1, \cdots, P_d]\,(d \ll D)$, the column vectors of the matrix P forms a basis spanning of the new space. By constraining the basis to be orthonormal, we have

$$\max_P \text{tr}\left(\mathbf{H}X^TPP^TX\mathbf{HL}\right) \quad \text{s.t.} \quad P_i^TP_j = \delta_{ij}. \quad (3)$$

To solve the problem, we have

$$\text{tr}\left(\mathbf{H}X^TPP^TX\mathbf{HL}\right) = \text{tr}\left(\sum_{i=1}^d \mathbf{H}X^TP_iP_i^TX\mathbf{HL}\right)$$
$$= \sum_{i=1}^d \text{tr}\left(\mathbf{H}X^TP_iP_i^TX\mathbf{HL}\right) = \sum_{i=1}^d P_i^T\left(X\mathbf{HLH}X^T\right)P_i \quad (4)$$

The optimal P_i^*'s $(1 \le i \le d)$ can be obtained easily by the *Lagrangian* method. If the eigenvalues of $X\mathbf{HLH}X^T$ are sorted as $\lambda_1 \ge \cdots \ge \lambda_D$, the optimal P_i^*'s are the normalized eigenvectors corresponding to the largest d eigenvalues. Since $X\mathbf{HLH}X^T$ is symmetric, the eigenvalues are all real. If the optimal projection P^* has been obtained, the corresponding HSIC value is $\sum_{i=1}^d \lambda_i$. Since the eigenvalues reflect the contribution of the corresponding dimensions, we can control d by setting a threshold thr $(0 \le thr \le 1)$ and then choosing the first d eigenvectors such that $\sum_{i=1}^d \lambda_i \ge thr \times (\sum_{i=1}^r \lambda_i)$. Thus, the optimization problem reduces to deriving eigenvalues of a $D \times D$ matrix and the computational complexity is $O(D^3)$. The Pseudo-code of the MDDM method is shown in Figure 1.

In the above analysis, we use inner product as kernel function on $\mathcal{Y}$, i.e., $l(\boldsymbol{y}, \boldsymbol{y}') = \langle\boldsymbol{y}, \boldsymbol{y}'\rangle$. If such a simple linear kernel is insufficient to capture the correlation between concepts, we can use a more delicate kernel function, e.g. quadratic or RBF. If L_{ij} can encode the correlation between labels $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, the dimensionality reduction process can get a better result with the guidance of $\mathbf{L}$.

MDDM(X, Y, d or thr)

Input:

X : $D \times N$ feature matrix
Y : $Q \times N$ label matrix
d : the dimensionality to be reduced to
thr: a threshold

Process:

1 Construct the label kernel matrix $\mathbf{L}$
2 Compute $X\mathbf{HLH}X^T$
3 **if** d is given
4 Do eigenvalue decomposition on $X\mathbf{HLH}X^T$, then construct $D \times d$ matrix P whose columns are composed by the eigenvectors corresponding to the largest d eigenvalues
5 **else** (i.e., thr is given)
6 Construct $D \times r$ matrix $\tilde{P}$ in a way similar to Step 4 where r is the rank of $\mathbf{L}$, then choose the first d eigenvectors that enable $\sum_{i=1}^d \lambda_i \ge thr \times \left(\sum_{i=1}^r \lambda_i\right)$ to compose P
7 **end if**

Output:

P: the projection from $\mathbb{R}^D$ to $\mathbb{R}^d$

Figure 1: Pseudo-code of the MDDM method

In contrast to previous HSIC methods which greedily selected a subset of features (Song *et al.* 2007) or using gradient descent to find a local optimal under the constraint of distance preserving (Song *et al.* 2008), we have a close-form solution for our purpose, which is effective and efficient.

Note that HSIC is just one among the many choices we can take to measure the dependence. We have also evaluated canonical component analysis (Hardoon, Szedmak, & Shawe-Taylor 2004) yet the speed is slower than the current MDDM. So we only report the results with HSIC here.

Experiments

We compare MDDM with three methods, including the linear dimensionality reduction method PCA (Jolliffe 1986), nonlinear dimensionality reduction method LPP (He & Niyogi 2004), and the only available multi-label dimensionality reduction method MLSI (Yu, Yu, & Tresp 2005). The multi-label k-nearest neighbor method ML-kNN with default setting $k = 10$ (Zhang & Zhou 2007) is used for classification after dimensionality reduction. As a baseline, we also evaluate the performance of ML-kNN in the original feature space (denoted by ORI). For LPP, the number of nearest neighbors used for constructing *adjacency graph* is as the same as that used in ML-kNN for classification. For MLSI, the parameter β is set to 0.5 as recommended in (Yu, Yu, & Tresp 2005). In the first series of experiments, the dimensionality of the lower-dimensional space, d, is decided by setting $thr = 99\%$. All dimensionality reduction methods reduce to the same dimensionality. The performance under different d values will be reported later in this section. For MDDM, we have evaluated different $l(\cdot, \cdot)$ (linear, quadratic and RBF) while the results are similar. So, here we only report the simplest case, i.e., the linear kernel.

Table 1: Average results (mean±std.) on 11 Yahoo data sets (↓ indicates "the smaller the better"; ↑ indicates "the larger the better")

Criterion	MDDM	MLSI	PCA	LPP	ORI
Hamming Loss ($\times 10^{-1}$) ↓	**0.394±0.134**	0.499±0.172	0.426±0.142	0.437±0.144	0.432±0.145
One-error ↓	**0.415±0.135**	0.539±0.066	0.469±0.155	0.488±0.161	0.471±0.157
Coverage ($\times 10$) ↓	**0.381±0.116**	0.888±0.223	0.417±0.129	0.433±0.131	0.410±0.124
Ranking Loss ↓	**0.092±0.038**	0.247±0.091	0.104±0.042	0.109±0.045	0.102±0.045
Average Precision ↑	**0.665±0.103**	0.489±0.061	0.624±0.115	0.607±0.119	0.625±0.116

We evaluate the performance of the compared methods using five criteria which are popularly used in multi-label learning, i.e., *hamming loss*, *one-error*, *coverage*, *ranking loss* and *average precision*. These criteria evaluate multi-label learning methods from different aspects, and it is difficult for one method to be better than another over all the criteria. Details of these criteria can be found in (Zhang & Zhou 2007).

Eleven web page classification data sets [1] are used in our experiments. The web pages were collected from the "yahoo.com" domain. Each data set corresponds to a top-level category of Yahoo. The web pages are classified into a number of second-level subcategories, and thus, one web page may belong to several subcategories simultaneously. Details of these data sets can be found in (Zhang & Zhou 2007).

The average results are shown in Table 1 where the best result on each evaluation criterion is highlighted in boldface. It is impressive that, pairwise t-tests at 95% significance level reveal that MDDM is significantly better than all the other methods on all the evaluation criteria.

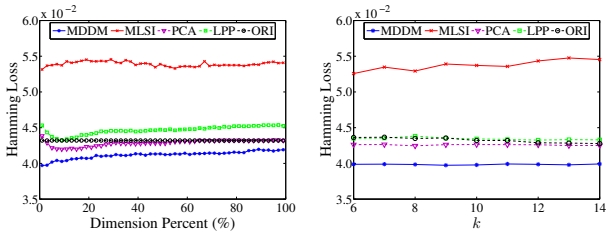
We also study the performance of the compared methods under different d, i.e., the dimensionality of the lower-dimensional space. We run experiments with d from 2% to 100% of the original space's dimensionality, with 2% as interval. Due to the page limit, we only present the results on *Hamming loss* which is arguably the most important multi-label evaluation criterion. It can be found from Figure 2(a) that the performance of MDDM with any d value is better than the best performance of the compared methods with their optimal d values. It is clear that MDDM is superior to the compared methods.

It is interesting to study that whether MDDM can also work well with different settings of k. So, we run experiments with k values ranging from 6 to 10 under the same d as that used in the first series of experiments. The results measured by *Hamming loss* are shown in Figure 2(b). It is nice to see that the performance of MDDM is quite robust to the setting of k, always better than the other methods.

Conclusion

In this paper, we propose the MDDM method, which performs multi-label dimensionality reduction by maximizing the dependence between the feature description and the associated class labels. It is easy to design variants of MDDM by using dependence measures other than HSIC to guide the induction of the lower-dimensional space. The label matrix $\mathbf{L}$ encoding the label correlation plays an important role in MDDM. Designing a better method for constructing $\mathbf{L}$ is an important future work.

[1] http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar.gz



(a) Different dimensionalities (b) Different k values

Figure 2: Average results on eleven Yahoo data sets

Acknowledgements We what to thank Kai Yu for providing the code of MLSI, and De-Chuan Zhan and Yang Yu for helpful discussion.

References

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2):179–188.

Gretton, A.; Bousquet, O.; Smola, A. J.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 63–77.

Hardoon, D. R.; Szedmak, S.; and Shawe-Taylor, J. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16(12):2639–2664.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *NIPS 16*.

Jolliffe, I. T. 1986. *Principal Component Analysis*. Berlin: Springer.

Song, L.; Smola, A.; Gretton, A.; Borgwardt, K.; and Bedo, J. 2007. Supervised feature selection via dependence estimation. In *ICML*, 823–830.

Song, L.; Smola, A.; Borgwardt, K.; and Gretton, A. 2008. Colored maximum variance unfolding. In *NIPS 20*. 1385–1392.

Yu, K.; Yu, S.; and Tresp, V. 2005. Multi-label informed latent semantic indexing. In *SIGIR*, 258–265.

Zhang, M.-L., and Zhou, Z.-H. 2007. ML-kNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7):2038–2048.