Multi-Label Informed Latent Semantic Indexing

Shipeng Yu[1,2]

Joint work with Kai Yu[1] and Volker Tresp[1]

August 2005

[1]Siemens Corporate Technology
Department of Neural Computation

[2]University of Munich
Institute for Computer Science

Outline

- Motivation
- Latent Semantic Indexing
- Multi-label Informed Latent Semantic Indexing (MLSI)
- Experimental Results
- Conclusion and Future works

Motivation

We are dealing with high-dimensional data in information retrieval.

A typical text corpus has more than 10,000 features (words as features)!

- What are the problems?
  - Noisy features: Effective features are small
  - Learnability: “curse of dimensionality”
  - Inefficiency: Computational cost is too high
- How to solve these problems? Dimensionality Reduction
  - Feature selection: Select part of the features
  - Latent semantic indexing (LSI): Learn a feature transformation from high-dimensional input space to a low-dimensional latent space

Why MLSI

- LSI is unsupervised:
  - Unable to use prior knowledge or label information
  - The indexing is not necessarily related to classification tasks
- We want to have a feature transformation method that can
  - Incorporate label information elegantly
  - Derive both linear and non-linear mappings
  - Explore the dependency between multiple categories
- This leads to Multi-label informed Latent Semantic Indexing (MLSI).

Before We Start ...

Some notations:

- We have N documents
- Document i is denoted as $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^M$
- Output for the ith document is denoted as $\mathbf{y}_i \in \mathcal{Y} \subset \mathbb{R}^L$
- $\mathbf{X} \in \mathbb{R}^{N \times M}$, $\mathbf{Y} \in \mathbb{R}^{N \times L}$ contain the input and output data as follows:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \vdots & \vdots \\ x_{N1} & \cdots & x_{NM} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1L} \\ \vdots & \vdots & \vdots \\ y_{N1} & \cdots & y_{NL} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_N^T \end{bmatrix}$$

- We aim to derive a mapping $\Psi : \mathcal{X} \mapsto \mathcal{V}$ such that $\mathcal{V} \subset \mathbb{R}^K$, $K < M$

Outline

- Motivation
- Latent Semantic Indexing
- Multi-label Informed Latent Semantic Indexing
- Experimental Results
- Conclusion and Future works

Latent Semantic Indexing

LSI finds the best rank-K approximation to the data matrix $\mathbf{X}$.

This can be equivalently solved by singular value decomposition (SVD) of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T$$

- We can sort diagonal entries of $\boldsymbol{\Sigma}$ in decreasing order
- $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$ gives the K mapping directions

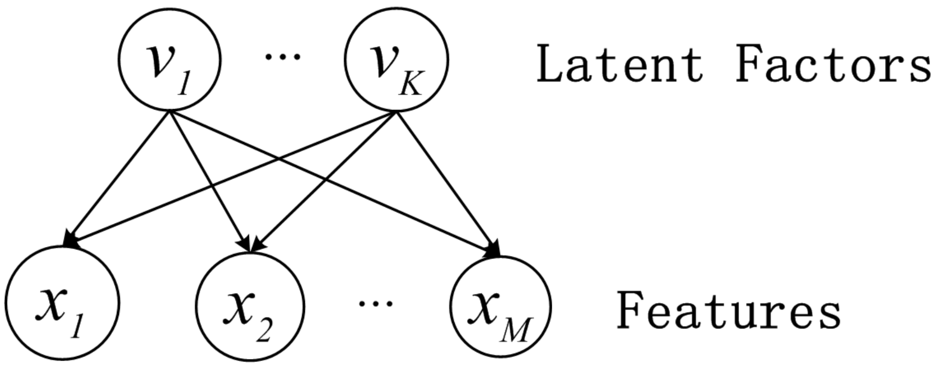
Problem: How to incorporate label information into the mappings?

Optimization Problem of LSI

Alternatively, LSI minimizes the reconstruction error of input data:

$$\min_{\mathbf{A},\mathbf{V}} \quad \|\mathbf{X} - \mathbf{V}\mathbf{A}\|_F^2$$
$$\text{s.t.} \quad \mathbf{V}^T\mathbf{V} = \mathbf{I},$$

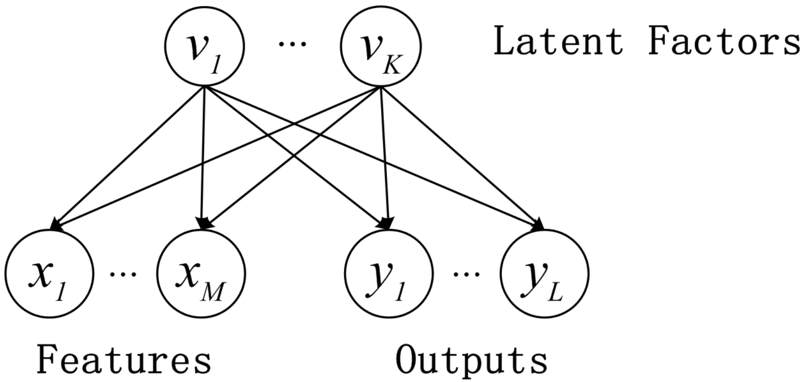
with $\mathbf{V} \in \mathbb{R}^{N\times K}$ the latent factors, and $\mathbf{A} \in \mathbb{R}^{K\times M}$ the factor loadings.



MLSI

In MLSI we are minimizing the reconstruction errors of both $\mathbf{X}$ and $\mathbf{Y}$:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{V}} \quad (1-\beta)\|\mathbf{X}-\mathbf{VA}\|_F^2 + \beta\|\mathbf{Y}-\mathbf{VB}\|_F^2$$
$$\text{s.t.} \quad \mathbf{V}^T\mathbf{V}=\mathbf{I}, \mathbf{V}=\mathbf{XW}.$$



- MLSI is biased by the outputs $\mathbf{Y}$
- MLSI minimizes the inter-correlation between $\mathbf{X}$ and $\mathbf{Y}$
- MLSI minimizes the intra-correlation within $\mathbf{Y}$ (if multiple outputs)

Outline

- Motivation
- Latent Semantic Indexing
- Multi-label Informed Latent Semantic Indexing
  - Primal form: Linear mappings
  - Dual form: Non-linear mappings
- Experimental Results
- Conclusion and Future works

Solution of MLSI

The optimization problem is

$$\min_{\mathbf{A},\mathbf{B},\mathbf{V}} \quad (1-\beta)\|\mathbf{X}-\mathbf{V}\mathbf{A}\|_F^2 + \beta\|\mathbf{Y}-\mathbf{V}\mathbf{B}\|_F^2$$
$$\text{s.t.} \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}, \mathbf{V} = \mathbf{X}\mathbf{W}.$$

Following standard Lagrange formulism, we obtain, at the optimum,

- $\mathbf{A}$ and $\mathbf{B}$ solely depend on $\mathbf{V}$: $\mathbf{A} = \mathbf{V}^T\mathbf{X}$, $\mathbf{B} = \mathbf{V}^T\mathbf{Y}$.
- Denote $\mathbf{K} := (1-\beta)\mathbf{X}\mathbf{X}^T + \beta\mathbf{Y}\mathbf{Y}^T$, the minimum value is $\sum_{i=K+1}^{N} \lambda_i$.
- We only need to optimize $\mathbf{W}$ since $\mathbf{V} = \mathbf{X}\mathbf{W}$.

MLSI: Primal Form

Denote $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$, we turn to an equivalent problem w.r.t. $\mathbf{w}$:

$$\max_{\mathbf{w}\in\mathbb{R}^M} \quad \mathbf{w}^T\mathbf{X}^T\mathbf{K}\mathbf{X}\mathbf{w}$$
$$\text{s.t.} \quad \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} = 1.$$

This leads to the primal form of the MLSI solution:

- Calculate $\mathbf{K} = (1-\beta)\mathbf{X}\mathbf{X}^T + \beta\mathbf{Y}\mathbf{Y}^T$;
- Solve a generalized eigenvalue problem $\mathbf{X}^T\mathbf{K}\mathbf{X}\mathbf{w} = \lambda\mathbf{X}^T\mathbf{X}\mathbf{w}$, obtain eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$ with largest K eigenvalues $\lambda_1 \geq \ldots \geq \lambda_K$;
- Form mapping functions $\psi_j(\mathbf{x}) = \sqrt{\lambda_j}\mathbf{w}_j^T\mathbf{x}, j = 1, \ldots, K$, and finally $\Psi(\mathbf{x}) = [\psi_1(\mathbf{x}), \ldots, \psi_K(\mathbf{x})]^T$ defines the mapping Ψ.

MLSI recovers LSI when $\beta = 0$.

MLSI: Dual Form

Dual form is obtained by applying representer theorem and define dual variable $\boldsymbol{\alpha}$ as

$$\mathbf{w} = \mathbf{X}^T\boldsymbol{\alpha}.$$

This leads to the equivalent dual form with respect to $\boldsymbol{\alpha}$:

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \quad & \boldsymbol{\alpha}^T\mathbf{K}_x\mathbf{K}\mathbf{K}_x\boldsymbol{\alpha} \\ \text{s.t.} \quad & \boldsymbol{\alpha}^T\mathbf{K}_x^2\boldsymbol{\alpha} = 1. \end{aligned}$$

$\mathbf{K}_x = \mathbf{X}\mathbf{X}^T, \mathbf{K}_y = \mathbf{Y}\mathbf{Y}^T, \mathbf{K} = (1-\beta)\mathbf{K}_x + \beta\mathbf{K}_y.$

This is a simpler problem for $N < M$.

Primal versus Dual

Which form to choose in real world applications?

- Primal MLSI solves an $M \times M$ generalized eigenvalue problem
  - more efficient when $M < N$
  - can only learn a linear mapping for $\mathbf{X}$
- Dual MLSI solves an $N \times N$ generalized eigenvalue problem
  - more efficient when $N < M$ (usually true for text data)
  - can learn non-linear mappings using kernel trick

Connection to Related Work

MLSI is more general to other supervised projection methods.

- Fisher Discriminant Analysis (FDA)
  - Only deal with binary classification problem
  - Can only handle one output
- Canonical Correlation Analysis (CCA)
  - Only minimize the correlation between $\mathbf{X}$ and $\mathbf{Y}$
  - Ignore intrinsic correlations of both $\mathbf{X}$ and $\mathbf{Y}$
- Partial Least Square (PLS)
  - A penalized CCA
  - Can not generalize well to new data

Outline

- Motivation
- Latent Semantic Indexing
- Multi-label Informed Latent Semantic Indexing
- Experimental Results
- Conclusion and Future works

Experiment Setup

The Goal: Evaluate indexing methods for multi-label classification.

- Data sets
  - Reuters-21578: 1600 documents with 6076 words, 47 categories
  - RCV1: 3588 documents with 5496 words, 79 categories
- Preprocessing
  - Take categories with at least 50 documents
  - Pick up words that occur at least 5 times in documents
  - Use TFIDF features

Methodology

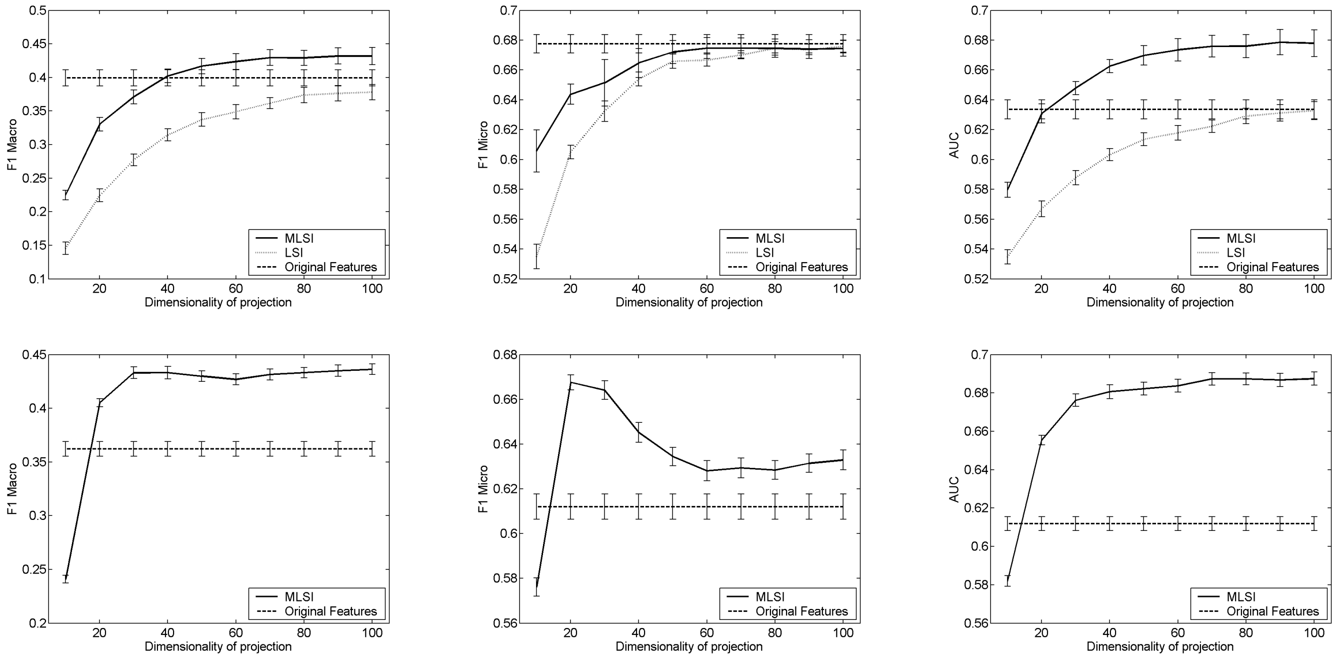
We compare three methods:

- Full Features: Use all features to do classification
- LSI: Classification with new unsupervised features
- MLSI: Classification with new supervised features

We test two settings for each data set:

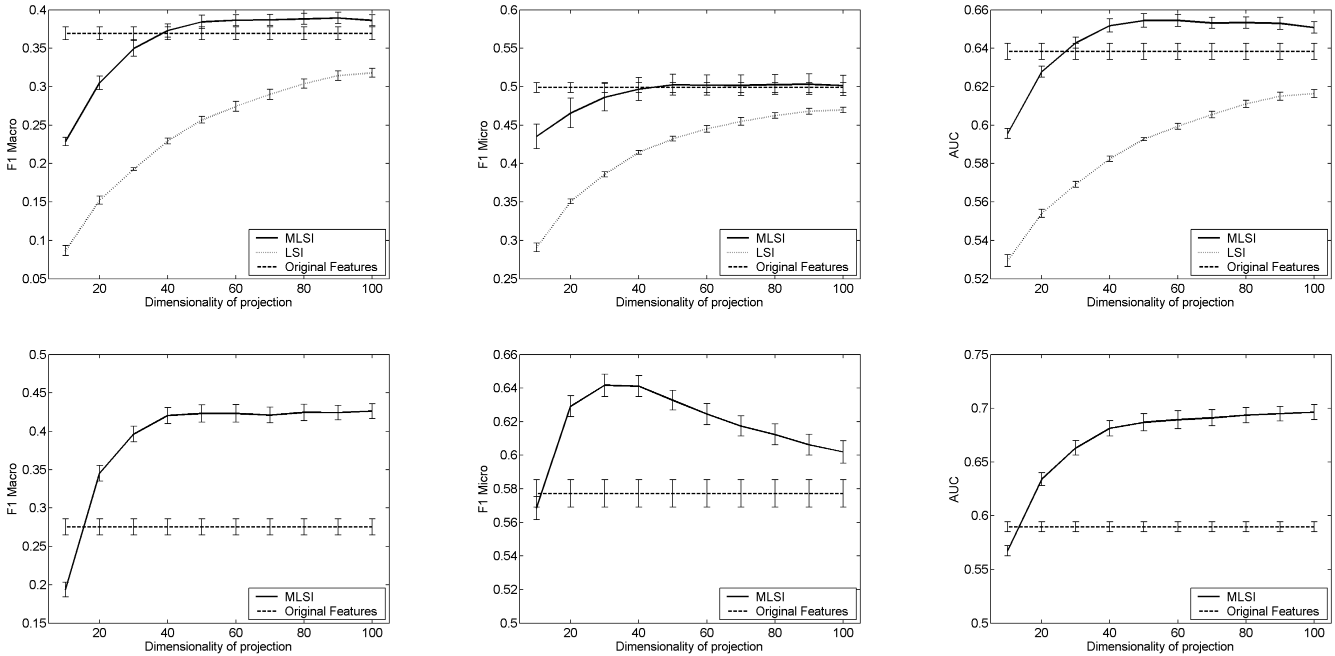
- Setting (I): We pick up 70% categories for classification and employ 5-fold cross-validation with one fold training and 4 folds testing
- Setting (II): Evaluate the classification performance on the rest 30% categories for previously unseen data with newly derived features

Results for Reuters-21578



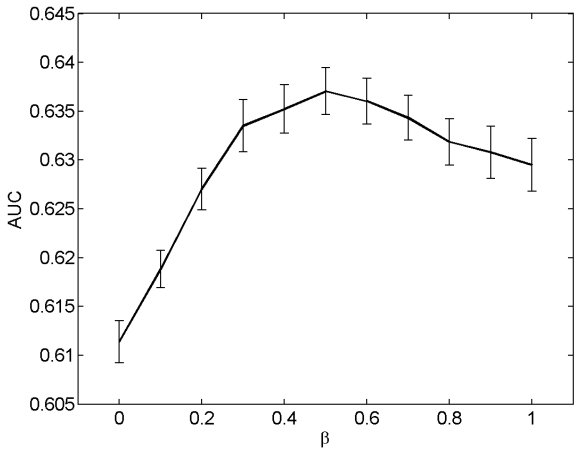
MLSI is significantly better than LSI.

Results for RCV1

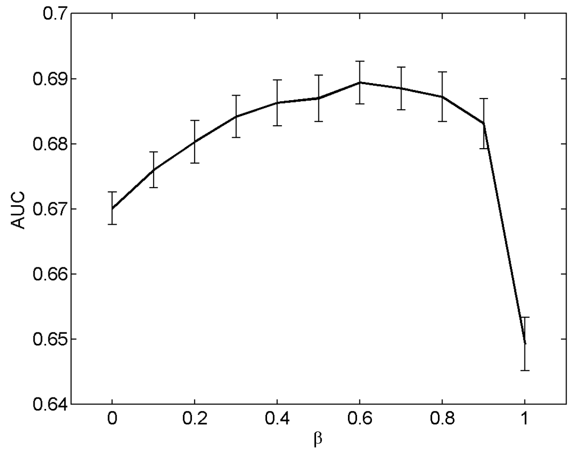


MLSI is significantly better than Full Features in setting (II).

Sensitivity of β for MLSI



setting (I)



setting (II)

Outline

- Motivation
- Latent Semantic Indexing
- Multi-label Informed Latent Semantic Indexing
- Experimental Results
- Conclusion and Future works

Conclusion

MLSI has the following advantages:

- It is supervised and incorporates label information
- It considers both the inter-correlation between $\mathbf{X}$ and $\mathbf{Y}$, and the intra-correlation of $\mathbf{Y}$
- Both linear and non-linear mappings are easy to derive
- It handles multiple outputs simultaneously
- It takes LSI as a special case (when $\beta = 0$)

Experimental results are very encouraging.

Future Works

- Compare with other supervised projection methods
- Automatically set parameter β
- Try larger data sets
- Apply the indexing to information retrieval tasks