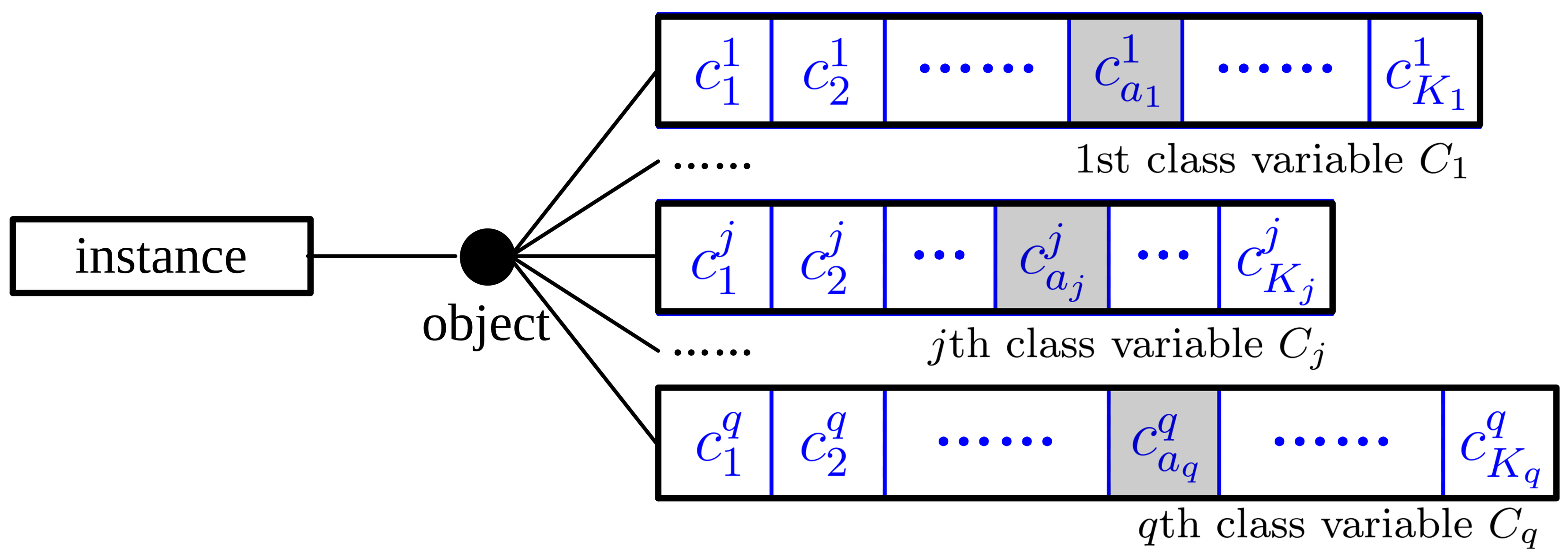
Maximum Margin Multi-Dimensional Classification(@AAAI'20)

贾彬彬† (jiabb@seu.edu.cn), 张敏灵§(zhangml@seu.edu.cn)

Introduction



Multi-Dimensional Classification (MDC)

Settings: $\mathcal{X} = \mathbb{R}^d$: d-dimensional input (feature) space

$\mathcal{Y} = C_1 \times C_2 \times \cdots \times C_q$: output space, where $C_j = \{c_1^j, c_2^j, \ldots, c_{K_j}^j\}$

Input : $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid 1 \le i \le N\}$: training data set, where

$\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^\top \in \mathcal{X}$ and $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iq}]^\top \in \mathcal{Y}$

Output : f : multi-dimensional classifier $\mathcal{X} \to \mathcal{Y}$

MDC example (A piece of music)

Dim. 1: Genre -->  --> rock, popular, classical, etc.

Dim. 2: Instrument --> --> piano, violin, guitar, etc.

Dim. 3: Language --> --> English, Chinese, Spanish, etc.

Our Goal : adapting maximum margin techniques for MDC

Two Key Challenges:

(I) modeling outputs from different dimensions are not comparable

(II) dependencies among different dimensions should be considered

The M³MDC Approach

S1: Transform $\mathcal{D}$ into $m = \sum_{j=1}^{q} \binom{K_j}{2}$ binary classification data sets via OvO decomposition w.r.t. each dimension;

S2: Solve the following maximum margin formulation:

$$\min_{\mathbf{W}, \boldsymbol{b}, \boldsymbol{\xi}, \mathbf{C}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \xi_j^i + \frac{\lambda_1}{2}\mathrm{tr}(\mathbf{W}\mathbf{W}^\top) + \frac{\lambda_2}{2}\mathrm{tr}(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^\top)$$

$$\text{s.t.} \quad y_j^i(\langle \boldsymbol{w}_i, \boldsymbol{x}_j^i \rangle + b_i) > 1 - \xi_j^i, \quad \mathbf{C} \succeq 0, \mathrm{tr}(\mathbf{C}) \le 1$$

$$\xi_j^i \ge 0, \ i = 1, \ldots, m, \ j = 1, \ldots, n_i$$

considering dependency

maximum margin

Notations: The ith OvO dataset: $\mathcal{D}^i = \{(\boldsymbol{x}_j^i, y_j^i) \mid 1 \le j \le n_i\}$ $(1 \le i \le m)$

The m hyperplanes: $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m] \in \mathbb{R}^{d \times m}$ and $\boldsymbol{b} = (b_1, \ldots, b_m)^\top$

$\mathbf{W}$'s column covariance matrix: $\mathbf{C} \in \mathbb{R}^{m \times m}$, regularization parameters: λ_1, λ_2

Remarks: (I) The optimization problem is jointly convex w.r.t. $\mathbf{W}$, $\boldsymbol{b}$ and $\mathbf{C}$;

(II) Due to the non-linear and non-smooth constraint $\mathbf{C} \succeq 0$, it is not easy to solve the optimization problem directly;

(III) In this paper, an alternating method is used to solve it:

```
repeat
    Optimizing with respect to W and b when C is fixed;
    Optimizing with respect to C when W and b are fixed;
until convergence
```

S3: Calculate m binary predictions $\boldsymbol{y}_*^b = \mathrm{sign}(\mathbf{W}^\top \boldsymbol{x}_* + \boldsymbol{b})$ for test instance $\boldsymbol{x}_*$;

S4: Return $\boldsymbol{x}_*$'s predicted class vector $\boldsymbol{y}_*$ via OvO decoding rule based on $\boldsymbol{y}_*^b$.

Optimization

☐**Optimizing with respect to W and *b* when C is fixed**

$$\min_{\mathbf{W}, \boldsymbol{b}, \boldsymbol{\xi}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \xi_j^i + \frac{\lambda_1}{2}\mathrm{tr}(\mathbf{W}\mathbf{W}^\top) + \frac{\lambda_2}{2}\mathrm{tr}(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^\top)$$

$$\text{s.t.} \quad y_j^i(\langle \boldsymbol{w}_i, \boldsymbol{x}_j^i \rangle + b_i) > 1 - \xi_j^i,$$

$$\xi_j^i \ge 0, \ i = 1, \ldots, m, \ j = 1, \ldots, n_i$$

Our Paper

The Lagrangian of the above problem is given by:

$$\mathcal{L}(\mathbf{W}, \boldsymbol{b}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \xi_j^i + \frac{\lambda_1}{2}\mathrm{tr}(\mathbf{W}\mathbf{W}^\top) + \frac{\lambda_2}{2}\mathrm{tr}(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^\top)$$

$$- \sum_{i=1}^{m} \sum_{j=1}^{n_i} \alpha_j^i [y_j^i(\langle \boldsymbol{w}_i, \boldsymbol{x}_j^i \rangle + b_i) - 1 + \xi_j^i] - \sum_{i=1}^{m} \sum_{j=1}^{n_i} \beta_j^i \xi_j^i$$

The dual problem, i.e., $\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{W}, \boldsymbol{b}} \mathcal{L}(\mathbf{W}, \mathbf{b})$, is equivalently formulated as:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i_1=1}^{m} \sum_{j_1=1}^{n_{i_1}} \sum_{i_2=1}^{m} \sum_{j_2=1}^{n_{i_2}} \alpha_{j_1}^{i_1} \alpha_{j_2}^{i_2} y_{j_1}^{i_1} y_{j_2}^{i_2} M_{i_1 i_2} \langle \boldsymbol{x}_{j_1}^{i_1}, \boldsymbol{x}_{j_2}^{i_2} \rangle - \sum_{i=1}^{m} \sum_{j=1}^{n_i} \alpha_j^i$$

$$\text{s.t.} \quad \sum_{j=1}^{n_i} \alpha_j^i y_j^i = 0 \ (1 \le i \le m), \ 0 \le \alpha_j^i \le 1 \qquad \boxed{\text{QP problem}}$$

where $\mathbf{M} = (\lambda_1 \mathbf{I}_m + \lambda_2 \mathbf{C})^{-\top} \mathbf{C}^\top$ and $\boldsymbol{\alpha} = (\alpha_1^1, \ldots, \alpha_{n_1}^1, \ldots, \alpha_1^m, \ldots, \alpha_{n_m}^m)^\top$.

After solving out $\boldsymbol{\alpha}$, we can obtain $\mathbf{W} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \alpha_j^i y_j^i \boldsymbol{x}_j^i \boldsymbol{e}_i^\top \mathbf{C} (\lambda_1 \mathbf{I}_m + \lambda_2 \mathbf{C})^{-1}$

and then obtain $\boldsymbol{b}$ via KKT conditions (for more details, please see our paper).

☐**Optimizing with respect to C when W and *b* are fixed**

$$\min_{\mathbf{C}} \ \mathrm{tr}(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^\top) \quad \text{s.t.} \ \mathbf{C} \succeq 0, \mathrm{tr}(\mathbf{C}) \le 1 \quad \Longrightarrow \quad \mathbf{C} = \frac{(\mathbf{W}^\top \mathbf{W})^{\frac{1}{2}}}{\mathrm{tr}((\mathbf{W}^\top \mathbf{W})^{\frac{1}{2}})}$$

closed-form solution

Experiments

Experimental Setup

- **10 Data sets** and **3 Evaluation metrics**
- **Comparing Algorithms:** BR,ECC,ECP,ESC [Read et al., TKDE14]
- **Experimental Protocol:** Ten-fold cross-validation

Experimental Results

Wilcoxon signed-ranks test for M³MDC against BR,ECC,ECP,ESC in terms of each evaluation metric (significance level $\alpha = 0.05$; p-values shown in the brackets).

Evaluation Metric	M³MDC vs BR	M³MDC vs ECC	M³MDC vs ECP	M³MDC vs ESC
Hamming Score	**win** [1.95e-3]	**win** [9.77e-3]	**win** [1.95e-3]	**win** [3.91e-3]
Exact Match	**win** [7.81e-3]	**tie** [7.70e-1]	**tie** [4.32e-1]	**tie** [7.54e-1]
Sub-Exact Match	**win** [2.34e-2]	**tie** [9.77e-2]	**win** [4.88e-2]	**tie** [1.95e-1]

☐Across all the 30 cases (10 data sets × 3 metrics), M³MDC ranks first in 21 cases;

☐In terms of *Hamming Score*, M³MDC is statistically better than BR/ECC/ECP/ESC;

☐In terms of all evaluation metrics, M³MDC is statistically better than BR;

☐For more details about experimental results and some further analysis (parameter sensitivity, correlation analysis, convergent characteristics), please see our paper.

Conclusion

A first attempt towards adapting maximum margin technique for MDC is investigated. Specifically, a novel approach named M³MDC is proposed which considers the margin over MDC examples via OvO decomposition and models the dependencies among class spaces with covariance regularization.

School of Computer Science and Engineering, Southeast University, China†, §
College of Electrical and Information Engineering, Lanzhou University of Technology, China†
Key Laboratory of Computer Network and Information Integration, Ministry of Education, China †, §
Collaborative Innovation Center of Wireless Communications Technology, China §