Supplementary Material for "Deep Multi-dimensional Classification with Pairwise Dimension-Specific Features"

To facilitate understanding, Table 1 summarizes the notations used in Section 3 and Algorithm 1 further presents the pseudo code of the proposed PIST approach. To eliminate the impact exerted by difference of base classifiers, we investigate five adjusted approaches including BR, KRAM, LEFA, MDKNN and SEEM by replacing the multi-class classifier with neural networks as PIST. In the context below, these approaches with changed base classifiers are denoted by the original name plus a subscript δ (e.g., BR_δ) and detailed experimental results are reported in Table 2. To show whether PIST achieves significantly superior/inferior performance against other comparing approaches on each data set, pairwise t-test at 0.05 significance level is conducted, where the corresponding win/tie/loss counts have been reported in Table 4 of the main body of this paper.

Table 1: Summary of the notations used in our paper.

Notation	Descriptions
d	number of features in input space
q	number of class spaces (dimensions) in output space
K_j	number of class labels in the j-th class space $(1 \leq j \leq q)$
m	number of MDC training examples
$\mathcal{X}$	the d-dimensional input (feature) space, i.e., $\mathcal{X} = \mathbb{R}^d$
C_j	the j-th class space where $C_j = \{c_1^j, c_2^j, \dots, c_{K_j}^j\}$ $(1 \leq j \leq q)$
c_a^j	the a-th class label in C_j $(1 \leq a \leq K_j)$
$\mathcal{Y}$	the output space where $\mathcal{Y} = C_1 \times C_2 \times \ldots \times C_q$
$\mathcal{D}$	the set of MDC samples where $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid 1 \leq i \leq m\}$
$\boldsymbol{x}_i$	the i-th feature vector where $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \dots, x_{id}]^\top \in \mathcal{X}$
$\boldsymbol{y}_i$	the i-th class vector where $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \dots, y_{iq}]^\top \in \mathcal{Y}$
$\phi(\boldsymbol{x})$	the feature embedding of feature vector $\boldsymbol{x}$.
f	the MDC predictive model: $\mathcal{X} \mapsto \mathcal{Y}$
$\boldsymbol{l}_{a_1a_2}$	a latent label embedding vector related to $c_{a_1}^1$ and $c_{a_2}^2$.
$\boldsymbol{l}^{(12)}$	the pairwise dimension embedding w.r.t. C_1 and C_2.
$f_{a_1a_2}$	the number of examples labeled by $c_{a_1}^1$ and $c_{a_2}^2$ in the training set.
$\bar{\boldsymbol{l}}_{a_1}$	the intra-class mean of $\{\boldsymbol{l}_{a_11}, \boldsymbol{l}_{a_12}, \dots, \boldsymbol{l}_{a_1K_2}\}(a_1 \in \{1, 2, \dots, K_1\})$.
$\bar{\boldsymbol{l}}'_{a_2}$	the intra-class mean of $\{\boldsymbol{l}_{1a_2}, \boldsymbol{l}_{2a_2}, \dots, \boldsymbol{l}_{K_1a_2}\}(a_2 \in \{1, 2, \dots, K_2\})$.
$\bar{\boldsymbol{l}}$	the global mean of all label embeddings w.r.t. C_1 and C_2.
σ	ReLU activate function.
$\boldsymbol{\theta}^{(12)}$	feature importance vector w.r.t. C_1 and C_2.
$\mathbf{W}_l, \boldsymbol{b}_l$	weight matrix and bias of the attention network.
$\mathbf{W}_s, \boldsymbol{b}_s$	weight matrix and bias of the network for transforming dimension-specific features.
$\mathbf{W}_o, \boldsymbol{b}_o$	weight matrix and bias of the softmax regression.
$\psi(x, y)$	the injective function from the Cartesian product $\{1, 2, \dots, K_1\} \times \{1, 2, \dots, K_2\} \to \{1, 2, \dots, K_1K_2\}$.
$\boldsymbol{s}^{(12)}$	pairwise dimension-specific features.
$\hat{\boldsymbol{p}}^{(12)}$	output vector of the softmax regression w.r.t. C_1 and C_2.
$\rho_u^{(r)}$	the predicted sum of probabilities that label is exactly c_u^r in the r-th dimension.
Q_u^r	normalized $\rho_u^{(r)}$ with softmax operation.
$\mathcal{L}_{ce}$	cross-entropy loss.
$\mathcal{L}_{le}$	label embedding loss.

Algorithm 1 The PIST approach

Input: MDC training set $\mathcal{D}$, an unseen instance $\boldsymbol{x}_*$, the dimension of label embeddings t

Output: Predicted class vector $\hat{\boldsymbol{y}}_*$ for $\boldsymbol{x}_*$

repeat
  Randomly sample an example $(\boldsymbol{x}, \boldsymbol{y})$ from $\mathcal{D}$
  for $u = 1$ to q **do**
    for $v = u + 1$ to q **do**
      Initialize label embeddings $\{\boldsymbol{l}_{a_1 a_2} | 1 \leq a_1 \leq K_u, 1 \leq a_2 \leq K_v\}$ by standard normal distribution and calculate $f_{a_1 a_2}$ according to its definition
      Obtain pairwise dimension embedding $\boldsymbol{l}^{(uv)}$ by Eq.(1)
      Calculate the label embedding loss $\mathcal{L}_{le}^{(uv)}$ by Eq.(4)
      Calculate pairwise dimension-specific feature $\boldsymbol{s}^{(uv)}$ by Eq.(5) and Eq.(6)
      Obtain predicted joint probability $\hat{\boldsymbol{p}}^{(uv)}$ by Eq.(10) and Eq.(11)
    end for
  end for
  for $j = 1$ to q **do**
    Initialize label embeddings $\{\boldsymbol{l}_a | 1 \leq a \leq K_j\}$ by standard normal distribution and calculate f_a according to its definition
    Obtain dimension embedding $\boldsymbol{l}^{(jj)}$ by Eq.(7)
    Calculate dimension-specific feature $\boldsymbol{s}^{(jj)}$ by Eq.(8) and Eq.(9)
    Obtain predicted joint probability $\hat{\boldsymbol{p}}^{(jj)}$ by Eq.(13) and Eq.(14)
    for $a = 1$ to K_j **do**
      Calculate normalized confidence score Q_a^j by Eq.(16) and Eq.(17)
    end for
    Obtain predicted result ω w.r.t. the j-th dimension by Eq.(20)
  end for
  Calculate the final loss $\mathcal{L}$ by Eq.(19)
  Update the trainable parameters with SGD algorithm
until *Converge*
Feed $\boldsymbol{x}_*$ to trained model by above steps and output predicted results $\hat{\boldsymbol{y}}_*$
Return $\hat{\boldsymbol{y}}_*$

Table 2: Experimental results (mean±std.) of MDC approaches of which base classifier are replaced with a multi-layer perceptron with a single hidden layer. In addition, •/◦ indicates whether Pist is significantly superior/inferior to other compared approaches on each data set with pairwise t-test at 0.05 significance level.

(a) Hamming Score

Data Set	Pist	Br_δ	Kram_δ	Lefa_δ	Mdknn_δ	Seem_δ
WQplants	.661±.013	.664±.014	.576±.023•	.662±.017	.615±.019•	.643±.038
WQanimals	.632±.014	.631±.010	.544±.020•	.635±.013	.566±.020•	.621±.017•
WaterQuality	.647±.012	.647±.009	.565±.011•	.648±.010	.588±.018•	.630±.017•
BeLaE	.452±.015	.411±.015•	.369±.017•	.406±.012•	.338±.019•	.382±.017•
Voice	.954±.008	.848±.019•	.963±.005◦	.834±.016•	.943±.006•	.553±.166•
Scm20d	.845±.012	.715±.010•	.844±.003	.754±.013•	.862±.004◦	.486±.048•
CoIL2000	.957±.004	.869±.004•	.954±.004•	.932±.006•	.871±.006•	.801±.099•
TIC2000	.945±.004	.931±.005•	.942±.008	.939±.003•	.862±.004•	.756±.153•
Flickr	.795±.003	.751±.005•	.752±.006•	.719±.008•	.726±.007•	.791±.005•
Adult	.725±.003	.722±.005	.672±.010•	.668±.007•	.685±.005•	.598±.078•
Default	.676±.003	.673±.003•	.659±.003•	.655±.002•	.652±.003•	.651±.045

(b) Exact Match

Data Set	Pist	Br_δ	Kram_δ	Lefa_δ	Mdknn_δ	Seem_δ
WQplants	.094±.021	.101±.029	.039±.019•	.100±.033	.068±.027•	.086±.033
WQanimals	.057±.015	.059±.021	.019±.013•	.062±.026	.026±.015•	.048±.019
WaterQuality	.009±.006	.008±.007	.002±.004•	.008±.005	.001±.003•	.007±.007
BeLaE	.035±.019	.025±.009	.006±.004•	.020±.011	.010±.010•	.016±.009•
Voice	.910±.016	.702±.038•	.928±.010◦	.677±.033•	.888±.013•	.291±.208•
Scm20d	.199±.019	.119±.006•	.185±.015•	.140±.009•	.224±.010◦	.000±.000•
CoIL2000	.822±.014	.450±.017•	.812±.016•	.716±.023•	.542±.018•	.274±.317•
TIC2000	.843±.013	.805±.015•	.834±.020	.825±.008•	.627±.011•	.335±.410•
Flickr	.330±.013	.237±.012•	.243±.017•	.195±.015•	.213±.011•	.317±.014•
Adult	.288±.006	.285±.011	.236±.012•	.222±.014•	.243±.007•	.127±.108•
Default	.195±.006	.192±.006	.174±.007•	.169±.005•	.174±.005•	.173±.044

(c) Sub-Exact Match

Data Set	Pist	Br_δ	Kram_δ	Lefa_δ	Mdknn_δ	Seem_δ
WQplants	.285±.050	.295±.042	.168±.044•	.299±.042	.216±.036•	.253±.074
WQanimals	.223±.042	.230±.032	.122±.040•	.237±.032	.141±.020•	.203±.040
WaterQuality	.053±.011	.050±.022	.011±.011•	.048±.025	.021±.016•	.046±.022
BeLaE	.160±.024	.124±.026•	.080±.021•	.125±.023•	.060±.017•	.088±.019•
Voice	.997±.003	.993±.005	.999±.002	.991±.003•	.997±.003	.816±.155•
Scm20d	.403±.025	.231±.014•	.382±.012•	.275±.014•	.452±.014◦	.001±.003•
CoIL2000	.966±.006	.898±.010•	.961±.006	.947±.008•	.863±.009•	.778±.174•
TIC2000	.993±.002	.990±.002•	.992±.004	.992±.003	.961±.004•	.936±.047•
Flickr	.723±.009	.620±.012•	.633±.016•	.554±.020•	.579±.014•	.716±.010
Adult	.693±.007	.684±.011•	.599±.017•	.595±.009•	.618±.007•	.472±.129•
Default	.610±.007	.605±.006•	.577±.007•	.571±.009•	.563±.007•	.564±.079