Domain Generalization via Invariant Feature Representation

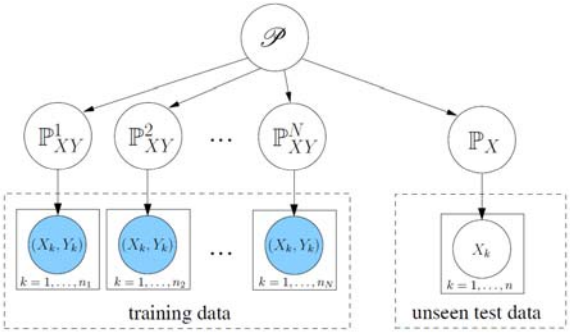
Krikamol Muandet; David Balduzzi, ; Bernhard Schölkopf
ICML, 2013

Angel Navia Vázquez
MLG/20-01-2014

Contenido

- Generalización de *dominio* (muestras no i.i.d.): ¿en qué problemas es aplicable?
- DICA: Domain Invariant Component Analysis
  - Varianza distribucional
  - El problema de optimización
  - Versión no supervisada
- Experimentos
- Conclusiones

Generalización de dominio (un tipo de "Transfer Learning")



- Las distribuciones marginales $\mathbb{P}_X$ suelen variar de un dominio a otro ("covariate shift": (Widmer 1996; Quiñonero 2009; Bickel 2009).
- Aprender un subespacio compartido es habitual en transfer learning (Argyriou 2007; Gu 2009; Passos 2012).
- La relación funcional o distribución condicional $\mathbb{P}_{Y|X}$ suele permanecer más estable o varía suavemente con el marginal $\mathbb{P}_X$.
- Los datos de test no son observados durante el entrenamiento (en "transfer learning" sí, lo que implica reentrenamiento para cada nuevo caso)

Nomenclatura

- $\mathcal{X}$ Espacio de entrada. $\mathcal{Y}$ Espacio de salida
- Dominio: distribución conjunta: $\mathbb{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$
- $\mathfrak{P}_{\mathcal{X}}$ and $\mathfrak{P}_{\mathcal{Y}|\mathcal{X}}$ conjuntos de distribuciones de probabilidad $\mathbb{P}_X$ on X and $\mathbb{P}_{Y|X}$ sobre $Y|X$
- N dominios observados a través de muestras (no i.i.d): $\mathcal{S} = \{S^i\}_{i=1}^N$, donde $S^i = \{(x_k^{(i)}, y_k^{(i)})\}_{k=1}^{n_i}$ viene de $\mathbb{P}_{XY}^i$
- $\mathcal{H}$ y $\mathcal{F}$ son RKHSs en $\mathcal{X}$ e $\mathcal{Y}$ con kernels y mappings inducidos:
  $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad x \to \phi(x) \in \mathcal{H}$
  $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R} \quad y \to \varphi(y) \in \mathcal{F}$
- Operadores de covarianza (se asume media cero):
  $\Sigma_{xx}, \Sigma_{yy}, \Sigma_{xy}$, and Σ_{yx}

Domain Invariant Component Analysis

- DICA: método kernel de transformación de los datos (los proyecta a un subespacio común) que:
  - A) minimiza la diferencia entre las distribuciones marginales $\mathbb{P}_X$
  - B) preserva las relaciones funcionales $\mathbb{P}_{Y|X}$ (generaliza bien en test)
  - C) extrae "invariantes": características comunes en múltiples dominios
- No requiere conocer la tarea objetivo (test), por tanto no requiere reentrenamiento para nuevas tareas
- DICA generaliza o está relacionado con algoritmos de reducción de dimensión bien conocidos: KPCA (Schölkopf 1998; Fukumizu 2004), TCA (Transfer Component Analysis, Pan 2011), o COIR (Covariance Operator Inverse Regression, Kim 2011)

Varianza distribucional

- Medida de disimilitud entre dominios
- Representamos distribuciones como elementos en un RKHS usando el "mean map" (kernel acotado y característico (P->∫ es inyectiva)):

$$\mu : \mathfrak{P}_{\mathcal{X}} \to \mathcal{H} : \mathbb{P} \mapsto \int_{\mathcal{X}} k(x, \cdot)\, \mathrm{d}\mathbb{P}(x) =: \mu_{\mathbb{P}}$$

- *NxN* matriz Gram de un conjunto de distrib. $\mathcal{P} = \{\mathbb{P}^1, \mathbb{P}^2 \ldots, \mathbb{P}^N\}$

$$G_{ij} := \langle \mu_{\mathbb{P}^i}, \mu_{\mathbb{P}^j} \rangle_{\mathcal{H}} = \iint k(x, z)\, \mathrm{d}\mathbb{P}^i(x)\, \mathrm{d}\mathbb{P}^j(z)$$

Definition 1. *Introduce probability distribution $\mathcal{P}$ on $\mathcal{H}$ with $\mathcal{P}(\mu_{\mathbb{P}^i}) = \frac{1}{N}$ and center G to obtain the covariance operator of $\mathcal{P}$, denoted as $\Sigma := G - \mathbf{1}_N G - G\mathbf{1}_N + \mathbf{1}_N G \mathbf{1}_N$. The* ***distributional variance*** *is*

$$\mathbb{V}_{\mathcal{H}}(\mathcal{P}) := \frac{1}{N}\mathrm{tr}(\Sigma) = \frac{1}{N}\mathrm{tr}(G) - \frac{1}{N^2}\sum_{i,j=1}^{N} G_{ij}. \quad (3)$$

Varianza distribucional Empírica

$$K = \begin{pmatrix} K_{1,1} & \cdots & K_{1,N} \\ \vdots & \ddots & \vdots \\ K_{N,1} & \cdots & K_{N,N} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

$$Q = \begin{pmatrix} Q_{1,1} & \cdots & Q_{1,N} \\ \vdots & \ddots & \vdots \\ Q_{N,1} & \cdots & Q_{N,N} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

$n = \sum_{i=1}^{N} n_i$ and $[K_{i,j}]_{k,l} = k(x_k^{(i)}, x_l^{(j)})$

$Q_{i,j} \in \mathbb{R}^{n_i \times n_j}$ equal $\begin{cases} (N-1)/(N^2 n_i^2) & \text{if } i = j \\ -1/(N^2 n_i n_j) & \text{otherwise} \end{cases}$

$$\widehat{\mathbb{V}}_{\mathcal{H}}(\mathcal{S}) = \frac{1}{N}\mathrm{tr}(\widehat{\Sigma}) = \mathrm{tr}(KQ)$$

Teoremas

Theorem 1. *Let* $\bar{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}^i$. *If* k *is a characteristic kernel, then* $\mathbb{V}_{\mathcal{H}}(\mathcal{P}) = \frac{1}{N}\sum_{i=1}^{N} \|\mu_{\mathbb{P}^i} - \mu_{\bar{\mathbb{P}}}\|_{\mathcal{H}}^2 = 0$ *if and only if* $\mathbb{P}^1 = \mathbb{P}^2 = \cdots = \mathbb{P}^N$.

(La varianza distribucional es válida como medida de divergencia ente dominios)

Theorem 2. *The empirical estimator* $\widehat{\mathbb{V}}_{\mathcal{H}}(\mathcal{S}) = \frac{1}{N}\mathrm{tr}(\widehat{\Sigma}) = \mathrm{tr}(KQ)$ *obtained from Gram matrix*

$$\widehat{G}_{ij} := \frac{1}{n_i \cdot n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} k(x_k^{(i)}, x_l^{(j)})$$

is a consistent estimator of $\mathbb{V}_{\mathcal{H}}(\mathcal{P})$.

Formulación de DICA

- Parámetros a ajustar: $B = [\beta_1, \beta_2, \ldots, \beta_m]$
- Encontrar una transformación $\mathcal{B}$ a un espacio de baja dimensión m<<n que **minimice la varianza distribucional** $\mathbb{V}_{\mathcal{H}}(\mathcal{S})$

(Reindexamos datos para más fácil manejo):

$$\{(x_k^{(i)}, y_k^{(i)})_{k=1}^{n_i}\}_{i=1}^{N} \Longrightarrow \{(x_k, y_k)\}_{k=1}^{n} \quad \text{where} \quad n = \sum_{i=1}^{N} n_i$$

$\Phi_x = [\phi(x_1), \phi(x_2), \ldots, \phi(x_n)]$

$\mathbf{b}_k = \sum_{i=1}^{n} \beta_k^i \phi(x_i) = \Phi_x \beta_k$ es la base k-ésima de $\mathcal{B}$ en $\mathcal{H}$

Proyección sobre la base k-ésima b_k

$$\tilde{\Phi}_x = \mathbf{b}_k^\top \Phi_x = \beta_k^\top \Phi_x^\top \Phi_x = \beta_k^\top K$$

Matriz de kernels en la proyección sobre $\mathcal{B}$

$$\widetilde{K} := \tilde{\Phi}_x^\top \tilde{\Phi}_x = KBB^\top K$$

Formulación de DICA (II)

- Varianza distribucional empírica:

$$\widehat{\mathbb{V}}_{\mathcal{H}}(\mathcal{BS}) = \mathrm{tr}(\widetilde{K}Q) = \mathrm{tr}(B^\top KQKB)$$

- Debe preservar la relación funcional entre X e Y. El **subespacio central C** es el mínimo subespacio que captura dicha relación.

Theorem 3. *If there exists a* ***central subspace*** $C = [\mathbf{c}_1, \ldots, \mathbf{c}_m]$ *satisfying* $Y \perp X | C^\top X$, *and for any* $a \in \mathbb{R}^d$, $\mathbb{E}[a^\top X | C^\top X]$ *is linear in* $\{\mathbf{c}_i^\top X\}_{i=1}^m$, *then* $\mathbb{E}[X|Y] \subset span\{\Sigma_{\mathrm{xx}} \mathbf{c}_i\}_{i=1}^m$.

-> bases de **C** coinciden con m mayores autovectores (Li 1991):

$$\mathbb{V}(\mathbb{E}[X|Y])\Sigma_{\mathrm{xx}}\mathbf{c} = \gamma \Sigma_{\mathrm{xx}}\mathbf{c}$$

Equiv. a resolver incrementalmente, preservando ortogonalidad:

$$\max_{\mathbf{c}_k \in \mathbb{R}^d} \frac{\mathbf{c}_k^\top \Sigma_{\mathrm{xx}}^{-1} \mathbb{V}(\mathbb{E}[X|Y])\Sigma_{\mathrm{xx}}\mathbf{c}_k}{\mathbf{c}_k^\top \mathbf{c}_k}$$

Formulación de DICA (III)

- Aproximación mediante operadores de covarianza (Kim 2011):

Theorem 4. *If for all $f \in \mathcal{H}$, there exists $g \in \mathcal{F}$ such that $\mathbb{E}[f(X)|y] = g(y)$ for almost every y, then*

$$\mathbb{V}(\mathbb{E}[X|Y]) = \Sigma_{\mathrm{xy}}\Sigma_{\mathrm{yy}}^{-1}\Sigma_{\mathrm{yx}} \ . \qquad (7)$$

- Aproximación empírica en Hilbert Space:

$$\Phi_y = [\varphi(y_1), \ldots, \varphi(y_n)] \qquad L = \Phi_y^\top \Phi_y$$

$$\widehat{\mathbb{V}}(\mathbb{E}[X|Y]) = \widehat{\Sigma}_{\mathrm{xy}}\widehat{\Sigma}_{\mathrm{yy}}^{-1}\widehat{\Sigma}_{\mathrm{yx}} = \tfrac{1}{n}\Phi_x L(L + n\varepsilon I_n)^{-1}\Phi_x^\top$$

$$\widehat{\Sigma}_{\mathrm{xy}} = \tfrac{1}{n}\Phi_x\Phi_y^\top \quad \widehat{\Sigma}_{\mathrm{yy}} = \tfrac{1}{n}\Phi_y\Phi_y^\top$$

donde ε es un regularizador del kernel

Formulación de DICA (IV)

- Suponiendo que existen las inversas de Σ_{yy} y Σ_{xy}:

$$\begin{aligned} \mathbf{b}_k^\top \widehat{\Sigma}_{\mathrm{xx}}^{-1}\widehat{\mathbb{V}}(\mathbb{E}[X|Y])\widehat{\Sigma}_{\mathrm{xx}}\mathbf{b}_k &= \frac{1}{n}\boldsymbol{\beta}_k^\top L(L + n\varepsilon I)^{-1}K^2\boldsymbol{\beta}_k \\ \mathbf{b}_k^\top \mathbf{b}_k &= \boldsymbol{\beta}_k^\top K\boldsymbol{\beta}_k, \qquad (8) \end{aligned}$$

- DICA: optimización conjunta en términos de B:

$$\max_{B\in\mathbb{R}^{n\times m}} \frac{\frac{1}{n}\mathrm{tr}\left(B^\top L(L + n\varepsilon I_n)^{-1}K^2B\right)}{\mathrm{tr}\left(B^\top KQKB + BKB\right)}$$

-> El numerador alinea B con las bases del subespacio central

-> El denominador controla disimilitud entre dominios y controla tamaño de B

Formulación de DICA (V)

- Reescribiendo como optimización con restricciones (Γ contiene los multiplicadores de Lagrange):

$$\mathcal{L} = \frac{1}{n}\mathrm{tr}\left(B^\top L(L+n\varepsilon I_n)^{-1}K^2B\right) - \mathrm{tr}\left(\left(B^\top KQKB + BKB - I_m\right)\Gamma\right)$$

- Derivando e igualando a cero obtenemos un problema generalizado de autovalores:

$$\frac{1}{n}L(L+n\varepsilon I_n)^{-1}K^2B = (KQK+K)B\Gamma$$

  - Evita dividir el espacio de salida ("slicing"), válido para alta dim. salida
  - Aplicable a salidas estructuradas (árboles, secuencias), donde "slicing" es imposible
  - Basado totalmente en kernels, aplicable a cualquier tipo de entrada o salida siempre que se pueda definir un kernel

Versión no supervisada (UDICA)

- Aplicable en dominios donde la salida no está disponible (image denoising, etc.)
- Caso especial de DICA:
  - $L = \frac{1}{n}I$ y ε->0
  - En vez de preservar el subespacio central maximiza la varianza de X

$$\max_{B\in\mathbb{R}^{n\times m}} \frac{\frac{1}{n}\mathrm{tr}(B^\top K^2B)}{\mathrm{tr}(B^\top KQKB + B^\top KB)}$$

- Calculable como problema generalizado de autovalores:

$$\frac{1}{n}K^2B = (KQK+K)B\Gamma$$

Resumen algoritmos DICA/UDICA

Algorithm 1 Domain-Invariant Component Analysis

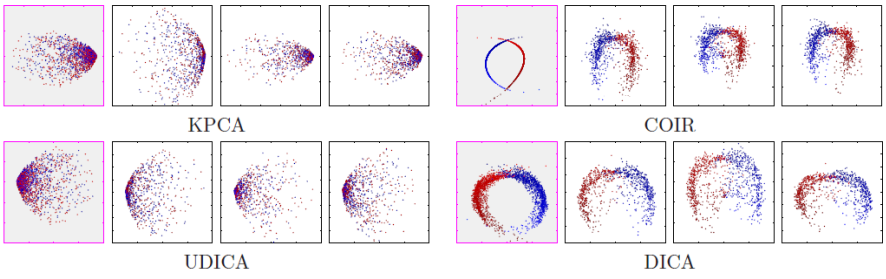
Input: Parameters λ, ε, and $m \ll n$.
Sample $\mathcal{S} = \{S^i = \{(x_k^{(i)}, y_k^{(i)})\}_{k=1}^{n_i}\}_{i=1}^{N}$.

Output: Projection $B_{n\times m}$ and kernel $\widetilde{K}_{n\times n}$.

1: Calculate gram matrix $[K_{ij}]_{kl} = k(x_k^{(i)}, x_l^{(j)})$ and $[L_{ij}]_{kl} = l(y_k^{(i)}, y_l^{(j)})$.
2: **Supervised:** $C = L(L + n\varepsilon I)^{-1}K^2$.
3: **Unsupervised:** $C = K^2$.
4: Solve $\frac{1}{n}CB = (KQK + K + \lambda I)B\Gamma$ for B.
5: Output B and $\widetilde{K} \leftarrow KBB^\top K$.
6: The test kernel $\widetilde{K}^t \leftarrow K^t BB^\top K$ where $K^t_{n_t\times n}$ is the joint kernel between test and training data.

Experimentos: Toy problem

- 10 colecciones de 200 puntos Poisson/Gauss 5d/Wishart
- Kernel Gauss con mismos parámetros en todos los casos
- En gris, datos de train, tres casos adicionales de test
- Color = valores de salida
- Posible sobreentrenamiento en COIR
- UDICA y DICA producen resultados más estables



Experimentos: Gating of Flow Cytometry Data

- Graft-versus-Host Disease (GvHD) dataset (injerto vs huésped)
- Muestras semanales de sangre periférica de 32 pacientes tras trasplante de médula (10 train, 20 test, 2 descartados)
- Objetivo: identificación de células CD3+CD4+CD8β+, muy correladas con la aparición de GvHD.
- Se espera encontrar un subespacio de células consistente con la variabilidad biológica entre pacientes e indicativas de GvHD
- Variantes SVM utilizadas:
  - Pooling SVM: junta todas las muestras de todos los pacientes
  - Distributional SVM: el kernel tiene en cuenta distribuciones de cada paciente (Blanchard 2011):

$$K(\tilde{x}_k^{(i)}, \tilde{x}_l^{(j)}) = k_1(\mathbb{P}^i, \mathbb{P}^j) \cdot k_2(x_k^{(i)}, x_l^{(j)})$$

Experimentos: Gating of Flow Cytometry Data (II)

- Acierto promedio en 30 particiones aleatorias

Methods	Pooling SVM			Distributional SVM		
	$n_i = 100$	$n_i = 500$	$n_i = 1000$	$n_i = 100$	$n_i = 500$	$n_i = 1000$
Input	91.68±.91	92.11±1.14	93.57±.77	91.53±.76	92.81±.93	92.41±.98
KPCA	91.65±.93	92.06±1.15	93.59±.77	**91.83±.60**	90.86±1.98	92.61±1.12
COIR	**91.71±.88**	92.00±1.05	92.57±.97	91.42±.95	91.54±1.14	92.61±.89
UDICA	91.20±.81	92.21±.19	93.02±.77	91.51±.79	91.74±1.08	93.02±.77
DICA	91.37±.91	**92.71±.82**	**94.16±.73**	91.51±.89	**93.42±.73**	**93.33±.86**

  - Con suficientes muestras DICA supera a los demás métodos
- Acierto LOO

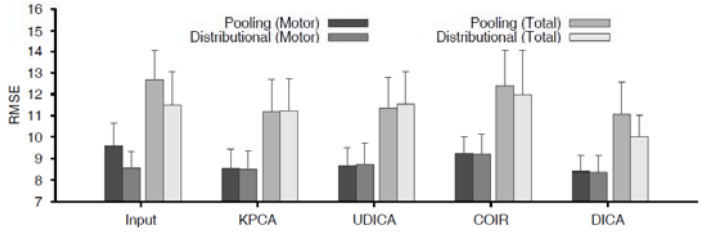
Methods	Pooling	Distributional
Input	92.03±8.21	93.19±7.20
KPCA	91.99±9.02	93.11±6.83
COIR	92.40±8.63	92.92±8.20
UDICA	92.51±5.09	92.74±5.01
DICA	**92.72±6.41**	**94.80±3.81**

  - Distributional mejor que Pooling
  - DICA mejor que el resto

Regresión: Telemonitorización de Parkinson

- Medida de voz de 42 pacientes (30 train) con Parkinson incipiente, 200 registros por paciente recogidos durante 6 meses
- Objetivo: predecir UPDRS (Unified Parkinson's Disease Rating Scale) total y motor
- Usan dos GPs (total y motor) $l(y_k^{(i)}, y_l^{(j)}) = \exp(-\|y_k^{(i)} - y_l^{(j)}\|^2/2\sigma_3^2)$
- DICA mejora pero no significancia estadística

Methods	Pooling GP Regression		Distributional GP Regression	
	motor score	total score	motor score	total score
LLS	8.82 ± 0.77	11.80 ± 1.54	8.82 ± 0.77	11.80 ± 1.54
Input	9.58 ± 1.06	12.67 ± 1.40	8.57 ± 0.77	11.50 ± 1.56
KPCA	8.54 ± 0.89	11.20 ± 1.47	8.50 ± 0.87	11.22 ± 1.49
UDICA	8.67 ± 0.83	11.36 ± 1.43	8.75 ± 0.97	11.55 ± 1.52
COIR	9.25 ± 0.75	12.41 ± 1.63	9.23 ± 0.90	11.97 ± 2.09
DICA	$\mathbf{8.40 \pm 0.76}$	$\mathbf{11.05 \pm 1.50}$	$\mathbf{8.35 \pm 0.82}$	$\mathbf{10.02 \pm 1.01}$



Conclusiones

- DICA aprende una transformación invariante de los datos que mejora la generalización inter-dominio
- *Teorema 5 sobre cotas* (no visto aquí): el error de generalización en dominios no vistos crece con la varianza distribucional
- DICA generaliza KPCA y COIR y está cercanamente relacionado con TCA
- En la práctica produce buenos resultados, sobre todo si se combina con SVM distribucional
- Válido para dominios en los que la relación funcional se mantiene estable en los distintos dominios
- Interesante posibilidad de optimizar conjuntamente características invariantes y clasificador si la varianza distribucional se introduce como regularizador en la función objetivo.

Referencias

- A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In Advances in Neural Information Processing Systems 19, pages 41-48. MIT Press, 2007.
- S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. Journal of Machine Learning Research, pages 2137-2155, 2009.
- G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In Advances in Neural Information Processing Systems 24, pages 2178-2186, 2011.
- K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel Dimensionality Reduction for Supervised Learning. In Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA, 2004.Q. Gu and J. Zhou. Learning the shared subspace for multi-task clustering and transductive transfer classification. In Proceedings of the 9th IEEE International Conference on Data Mining, pages 159-168. IEEE Computer Society, 2009.
- M. Kim and V. Pavlovic. Central subspace dimensionality reduction using covariance operators. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(4):657-670, 2011.
- K.-C. Li. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414):316-327, 1991.
- S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks, 22(2):199-210, 2011.
- A. Passos, P. Rai, J. Wainer, and H. D. III. Flexible modeling of latent task structures in multitask learning. In Proceedings of the 29th international conference on Machine learning, Edinburgh, UK, 2012.
- J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. Dataset Shift in Machine Learning. MIT Press, 2009.
- B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(5):1299-1319, July 1998.
- G. Widmer and M. Kurat. Learning in the Presence of Concept Drift and Hidden Contexts. Machine Learning, 23:69, 101, 1996.

Gracias!