Semi-Supervised Domain Adaptation with Non-Parametric Copulas

David Lopez-Paz[1], Jose Miguel Hernández-Lobato[2], Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems; [2]Cambridge University

Abstract

We address the problem of domain adaptation [4]. We focus on how to use data from a *source task* to help estimate another different but related *target task*, for which limited access to data is assumed.

Our approach consists in:

1. Modeling each of the tasks as a density estimation problem.
2. Estimating the *source task* as a Regular Vine (R-Vine) Copula Decomposition.
3. Detecting and adapting changes in each of the R-Vine factors accross tasks.

Also, we propose a new non-parametric Vine Copula Distribution. Finally, we analyze the performance of the proposed framework in real-world data and compare it to state-of-the-art algorithms.

Copulas

The copula $c(\mathbf{u})$ of a given density $p(\mathbf{x})$ describes the dependencies among the rr.vv. $x_1, \ldots, x_d$, but contains no information about their marginal effects, since $P_X(X) \sim U(0,1)$ for any r.v. X [6]:

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_i(x_i) \underbrace{c(P(x_1), ..., P(x_d))}_{\text{copula}}.$$

Given a sample $\{(x_i, y_i)\}_{i=1}^n$ from $p(x, y)$, we can obtain pseudo-samples from its copula c by mapping each observation to the unit square using marginal c.d.f.s, i.e.:

$$U := \{(u_i, v_i)\}_{i=1}^n := \{(\hat{P}(x_i), \hat{P}(y_i))\}_{i=1}^n.$$

Non-Parametric Bivariate Copulas

KDE using U is not possible due to its bounded support [2]. Instead, we back-transform the data to have Gaussian marginals using the Gaussian q.d.f. Φ^{-1}, and estimate c as:

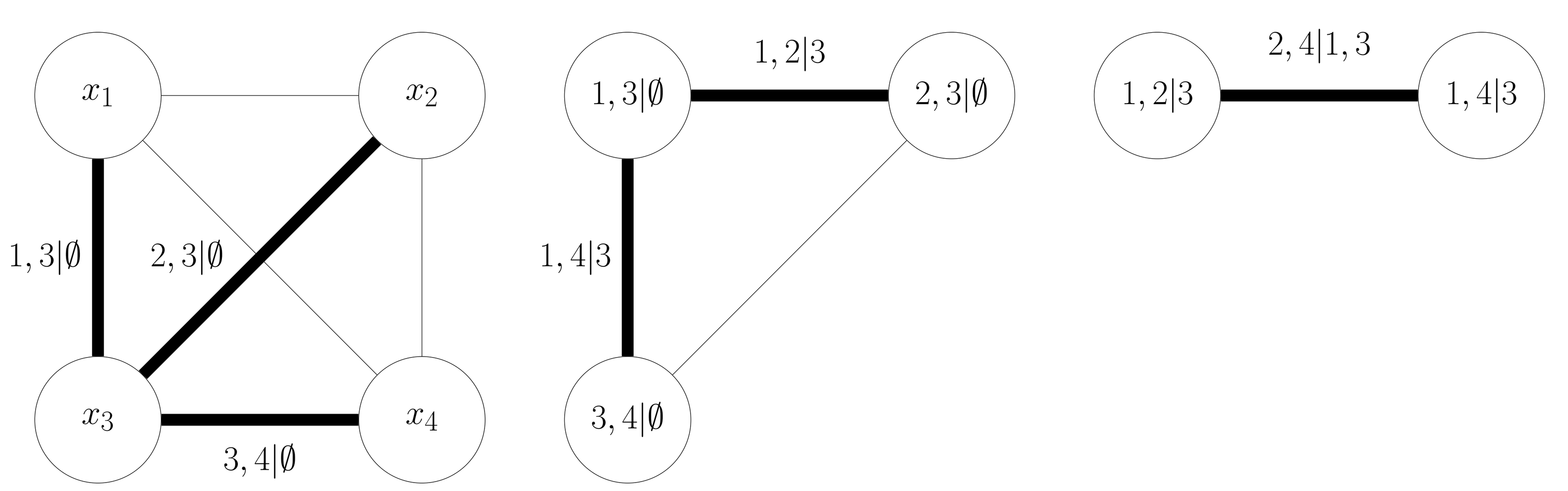
$$\hat{c}(u, v) = \frac{\hat{p}(\Phi^{-1}(u), \Phi^{-1}(v))}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathcal{N}(\Phi^{-1}(u), \Phi^{-1}(v)|\Phi^{-1}(u_i), \Phi^{-1}(v_i), \mathbf{\Sigma})}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))}. \quad (1)$$

Regular Vine Copula Distributions

High-dimensional copulas can be factorized as a collection of conditional bivariate copulas, in a so-called Regular Vine Copula Distribution or Decomposition [1]:

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_i(x_i) \prod_{i=1}^{d-1} \prod_{e(j,k)\in E_i} c_{jk|D(e)}(P_{j|D(e)}(x_{j|D(e)}), P_{k|D(e)}(x_{k|D(e)})) \quad (2)$$

These models are formed by a collection of trees, in which each edge represents a bivariate copula:



$$p_{1234} = \underbrace{p_1 \cdot p_2 \cdot p_3 \cdot p_4}_{\text{marginals}} \cdot \underbrace{c_{12|\emptyset} \cdot c_{13|\emptyset} \cdot c_{34|\emptyset}}_{\text{tree 1}} \cdot \underbrace{c_{23|1} \cdot c_{14|3}}_{\text{tree 2}} \cdot \underbrace{c_{24|13}}_{\text{tree 3}}$$

- Each bivariate copula forming the vine can belong to a different parametric family.
- The election of each spanning tree is done by maximizing the sum of the absolute empirical Kendall's τ's associated with each edge.
- Conditional c.d.f.s at tree i are expressed as partial derivatives of copulas from tree $i-1$ [1]:

$$P(u|\mathbf{v}) = \frac{\partial C_{u,v_j|\mathbf{v}_{-j}}(P_{u|\mathbf{v}_{-j}}(u|\mathbf{v}_{-j}), P_{v_j|\mathbf{v}_{-j}}(v_j|\mathbf{v}_{-j}))}{\partial P_{v_j|\mathbf{v}_{-j}}(v_j|\mathbf{v}_{-j})}. \quad (3)$$

Non-Parametric Regular Vine Copula Distributions

We now introduce the rule in (3) to build non-parametric Regular Vine Copula Distributions that are formed by non-parametric bivariate copula densities as the ones in (1):

$$\hat{P}(u|v) = \int_0^u \hat{c}(x, v)\,dx = \frac{1}{n\phi(w)} \sum_{i=1}^{n} \int_0^u \frac{\mathcal{N}(\Phi^{-1}(x), w|z_i, w_i, \mathbf{\Sigma})}{\phi(\Phi^{-1}(x))} dx =$$

$$= \frac{1}{n\phi(w)} \sum_{i=1}^{n} \mathcal{N}(w|w_i, \sigma_w^2)\, \Phi\left[\frac{\Phi^{-1}(u) - \mu_{z_i|w_i}}{\sigma^2_{z_i|w_i}}\right].$$

Domain Adaptation with Regular Vines

Assume we are given samples from two different but related tasks, namely the source task data:

$$D_s = \{\mathbf{x}_i, y_i\}_{i=1}^{N_s},$$

and the target task data:

$$D_t = \{\mathbf{x}'_i, y'_i\}_{i=1}^{N_t},$$

with $N_t << N_s$. Our objective is to use the maximum amount of information from the source task data to infer a better model for the target task data.

We exemplify the proposed technique for the two dimensional feature space case. First, we infer an R-Vine decomposition p_s from the source task data (equation 2), and clone it in a second R-Vine distribution p_t for the target task, namely:

$$p_s(\mathbf{x}, y) = p_1(x_1)\ p_2(x_2)\ p_y(y)\ c_{12}(P_1(x_1), P_2(x_2))\ c_{1y}(P_1(x_1), P_y(y))$$

$$p_t(\mathbf{x}, y) = p'_1(x_1)\ p'_2(x_2)\ p'_y(y)\ c'_{12}(P'_1(x_1), P'_2(x_2))\ c'_{1y'}(P'_1(x_1), P'_y(y))$$

When comparing pairs of factors using the target task data, several adaptation scenarios arise:

1. $p_i \neq p'_i$ or $P_i \neq P'_i$: p'_i and P'_i are reestimated using the corresponding samples from D_t.
2. $p_i = p'_i$ or $P_i = P'_i$: p'_i and P'_i are reestimated using the corresponding samples from $D_t \cup D_s$.
3. $c_{ij|k} \neq c_{ij|k}$: $c_{ij|k}$ is reestimated using the corresponding samples from D_t.
4. $c_{ij|k} = c_{ij|k}$: $c_{ij|k}$ is reestimated using the corresponding samples from $D_t \cup D_s$.

Several of the previous can occur, but there is no limitation in addressing them independently.

Detecting Changes

We propose the use of the *Maximum Mean Discrepancy* (MMD, [3]) test. Given samples $\mathbf{x}$ and $\mathbf{y}$ from two distributions X and Y, MMD will determine $X \overset{?}{=} Y$ according to distance between the embeddings of the empirical distributions of these two samples in a RKHS is significantly large. The empirical form of this statistic takes the form:

$$\text{MMD}(\mathbf{x}, \mathbf{y}) = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)$$

Semi-Supervised Domain Adaptation

Note that unlabeled, unpaired or incomplete data is still helpful to refine the factors of the R-Vine Decomposition p_t that depend on it. Please refer to the experiments.

Experimental Results

Average Test Log-Likelihood for Density Estimation:

Dataset	Auto	Cloud	Housing	Magic	Page-Blocks	Wireless
No. of variables	8	10	14	11	10	11
KDE	1.32 ± 0.06	3.25 ± 0.10	1.96 ± 0.17	1.13 ± 0.11	1.90 ± 0.13	0.98 ± 0.06
PRV	1.84 ± 0.08	**5.00 ± 0.12**	1.68 ± 0.11	2.09 ± 0.08	4.69 ± 0.20	0.36 ± 0.08
NPRV	**2.07 ± 0.07**	4.54 ± 0.13	**3.18 ± 0.17**	**2.72 ± 0.17**	**5.64 ± 0.14**	**2.17 ± 0.13**

NMSE for Domain Adaptation Regression:

Dataset	Wine	Sarcos	Rocks-Mines	Hill-Valleys	Axis-Slice	Isolet
No. of variables	12	21	60	100	386	617
GP-Source	0.86 ± 0.02	1.80 ± 0.04	0.90 ± 0.01	1.00 ± 0.00	1.52 ± 0.02	1.59 ± 0.02
GP-All	0.83 ± 0.03	1.69 ± 0.04	1.10 ± 0.08	0.87 ± 0.06	1.27 ± 0.07	1.58 ± 0.02
Daume	0.97 ± 0.03	0.88 ± 0.02	0.72 ± 0.09	0.99 ± 0.03	0.95 ± 0.02	0.99 ± 0.00
SSL-Daume	0.82 ± 0.05	0.74 ± 0.08	0.59 ± 0.07	0.82 ± 0.07	0.65 ± 0.04	0.64 ± 0.02
ATGP	0.86 ± 0.08	0.79 ± 0.07	**0.56 ± 0.10**	0.15 ± 0.07	1.00 ± 0.01	1.00 ± 0.00
KMM	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
KuLSIF	0.91 ± 0.08	1.67 ± 0.06	0.65 ± 0.10	0.80 ± 0.11	0.98 ± 0.07	0.58 ± 0.02
NPRV	**0.73 ± 0.07**	**0.61 ± 0.10**	0.72 ± 0.13	**0.15 ± 0.07**	0.38 ± 0.07	0.46 ± 0.09
UNPRV	0.76 ± 0.06	0.62 ± 0.13	0.72 ± 0.15	0.19 ± 0.09	**0.37 ± 0.07**	**0.42 ± 0.04**
Av. Ch. Mar.	10	1	38	100	226	89
Av. Ch. Cop.	5	8	49	34	155	474

Notes:

1. All features are continuous to ensure copula functional form uniqueness.
2. Our method is abbreviated as NPRV (Non-Parametric Regular Vine).
3. Experiments are based on 50 random 1000-sample training/test splits.
4. Only 5% from the target data task was labeled.
5. UNPRV ignores this so-said labeled data from the target task.

Future Work

We are investigating how to improve the proposed framework in the following directions:

1. Inclusion of discrete features.
2. More advanced techniques of factor refinement, beyond the presented substitution.
3. Domain Adaptation rules for parametric vines: correction of bivariate copula families or parameters accross domains.

References

[1] K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198, 2006.

[2] J. Fermanian and O. Scaillet. The estimation of copulas: Theory and practice. *Copulas: From Theory to Application in Finance*, pages 35–60, 2007.

[3] A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola. A kernel method for the two-sample-problem. *NIPS*, pages 513–520, 2007.

[4] S. Jialin-Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[5] H. Joe. Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. *Distributions with Fixed Marginals and Related Topics*, 1996.

[6] R. Nelsen. *An Introduction to Copulas.* Springer Series in Statistics, 2nd edition, 2006.