Transfer Learning with Copulas

David López Paz
MPI for Intelligent Systems
david.lopez@tuebingen.mpg.de

José Miguel Hernández-Lobato
University of Cambridge
jmh233@cam.ac.uk

Transfer learning methods use the knowledge gained by learning a particular *source* task to address more efficiently another different but related *target* task. They relax the frequent assumption in machine learning which states that both training and test data are drawn from the same distribution. This framework is most useful in those learning problems in which the amount of training data is reduced while, at the same time, some related tasks do have large amounts of available data. A recent survey of transfer learning techniques is described in [1].

Copulas allow to construct multivariate models by separating the modeling of marginals from the modeling of dependence. Under this framework, any form of dependence is represented by a unique copula function [2]. When two regression tasks share the same dependence structure, we may be able to transfer knowledge between them by means of a common copula model, allowing the marginal distributions to vary freely. In this manner, we reduce the problem of learning a multivariate model for the target task into the problem of learning i) the marginal distributions of the data and ii) a copula model which is shared by the two tasks.

This new approach is validated in a series of experiments with both simulated and real-world data. In these experiments, the target task has over 100 times less training instances than the source task. Results indicate that the transfer of knowledge by means of a common copula model is effectively accomplished. The proposed method is compared with standard support vector regression methods. The marginals are estimated in a non-parametric manner using kernels. For the estimation of the copula function, we use either parametric Gaussian copulas or non-parametric models based on kernels.

1 Transfer Learning by Means of a Common Copula Model

Consider two regression tasks, a source task s and target task t, with respective unidimensional inputs X_s and X_t and outputs Y_s and Y_t. We assume that the two tasks share the same dependence structure between their input and output variables, and hence, can be described by the same copula function. In addition, the amount of data for task s is supposed to be much larger than the amount of data for task t. Our objective is to identify the common copula function using the data from both tasks and use it to improve the performance of predictions in task t.

Given samples $\{(x_s^{(i)}, y_s^{(i)})\}_{i=1}^{n_s}$ and $\{(x_t^{(i)}, y_t^{(i)})\}_{i=1}^{n_t}$ from tasks s and t, respectively, we obtain a pseudo-sample from the common copula function by mapping each observation to the unit interval using the empirical estimates of the marginal distributions. In this manner, we obtain the sample $\{(\hat{F}_{X_s}(x_s^{(i)}), \hat{F}_{Y_s}(y_s^{(i)}))\}_{i=1}^{n_s} \cup \{(\hat{F}_{X_t}(x_t^{(i)}), \hat{F}_{Y_t}(y_t^{(i)}))\}_{i=1}^{n_t}$, where $\hat{F}_{X_s}$, $\hat{F}_{X_t}$, $\hat{F}_{Y_s}$ and $\hat{F}_{Y_t}$ are the estimated marginal distributions. The transformation of a random variable by the evaluation of its cumulative probability is known as the *probability integral transform* (PIT). This latter sample is used to obtain an estimate $\hat{c}$ of the copula density. Given $\hat{c}$, $\hat{F}_{X_t}$ and $\hat{F}_{Y_t}$, we can approximate the conditional density of Y_t given X_t as

$$\hat{p}(y_t|x_t) = Z^{-1}\hat{c}(\hat{F}_{X_t}(x_t), \hat{F}_{Y_t}(y_t))\hat{f}_{Y_t}(y_t)\,, \qquad (1)$$

where Z is a normalization constant and $\hat{f}_{Y_t}$ is the empirical marginal density of Y_t. For prediction, we compute the conditional mean using (1).

2 Results and Conclusions

In the following experiments, the source task has 1000 data points, while the target task contains only 10 observations. The cumulative distributions $\hat{F}_{X_t}$, $\hat{F}_{Y_t}$, $\hat{F}_{X_s}$, $\hat{F}_{Y_s}$, and the density $\hat{f}_{Y_t}$ are estimated using Gaussian kernels whose width is adjusted by Silverman's rule [4]. For the construction of the copula estimate $\hat{c}$ we use two approaches. The first one (G. Copula) is based on fitting a Gaussian copula by maximum likelihood. The second approach (K. Copula) does not make any assumption on the specific form of the dependence between the random variables, and employs kernels for the estimation of the copula density [3].

Simulated Data: The data for both regression problems are sampled from two bivariate Gaussian distributions with equal covariance matrix ($\sigma_x = \sigma_y = 1$, $\rho = 0.5$) but different mean vectors. In the target task, the mean components of the Gaussian distribution, that is, μ_{X_t} and μ_{Y_t}, are zero. By contrast, in the source task these components, μ_{X_s} and μ_{Y_s}, are equal to 10.

Real-world Data: In this case, the data for both regression problems is obtained from the UCI *Abalone* dataset [5]. X_s and X_t correspond to the *age* attribute, while Y_s and Y_t are the *length* and *diameter* attributes, respectively.

Dataset	Method	Avg. MSE
Simulated	**G. Copula**	**0.886±0.201**
	K. Copula	0.892±0.193
	SVM$_t$	1.112±0.297
	SVM$_{s+t}$	0.968±0.262
Abalone	G. Copula	0.011±0.023
	K. Copula	**0.008±0.001**
	SVM$_t$	0.015±0.004
	SVM$_{s+t}$	1.108±0.139

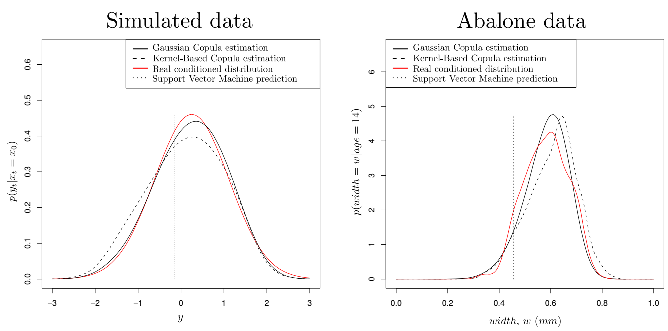

Figure 1: Left, average test MSE for each method on each problem. Middle and right, plots of $p(y_t|x_t)$, estimations of this function by the different methods and SVM prediction, for simulated data (middle, $x_t = 2$) and Abalone dataset (right, $x_t = 14$).

The table in the left-hand side of Figure 1 shows the average test MSE for each method on 100 train-test partitions of the data. The copula approach is compared with support regression methods (SVM) using a Gaussian kernel, trained with i) the target task data (SVM$_t$) and ii) the normalized union of the source and target task data (SVM$_{s+t}$). SVM parameters are fixed by cross-validation. The plots on the middle and right of Figure 1 show examples of $p(y_t|x_t)$ for a particular x_t value on each dataset. We include the true conditional density (red) and the estimates computed using the Gaussian copula (solid black line) or the kernel-based copula (dashed black line). The plots also show the prediction of SVM$_t$. Overall, the best performing method is the one based on copulas.

These experiments show that the proposed method effectively accomplishes the transference of learning between tasks. For future work, we plan to extend the method to higher dimensions and use several source tasks for transfer learning instead of only one.

References

[1] Jialin Pan S., Yang Q. (2010), *A Survey on Transfer Learning*, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10.

[2] Nelsen, Roger B. (1999), *An Introduction to Copulas*, New York: Springer.

[3] Fermanian, J., Scaillet, O. (2003), *Nonparametric estimation of copulas for time series*. The Journal of Risk 5 (4), 25–54.

[4] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Vol. 26 of Monographs on Statistics and Applied Probability, Chapman and Hall, London.

[5] UCI Abalone Data Set: `http://archive.ics.uci.edu/ml/datasets/Abalone`