ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1089

Transductive Learning for Multi-Task Copula Processes[1]

Markus Schneider[2] and **Fabio Ramos**[3]

1 Introduction & Related Work

We tackle the problem of multi-task learning with copula process. *Multi-task learning* is valuable in many areas of research such as spatial-temporal modeling, environmental sciences, numerical optimization and data fusion. In these problems it is advantageous to predict more than one quantity at a time (in contrast to single-task learning) to exploit inter-dependencies. Kernel-based algorithms achieve this by the use of an appropriate *multi-task* kernel. *Gaussian process* (GP) [6] based regression, as a simple and fully probabilistic model, is often the tool of choice for such problems. *Copulas*, with roots in statistics are models that separate the dependence structure of two or more random variables from their marginal distribution, thus possessing the flexibility of using a different probability distribution function for each variable. Copula distributions can be extended to stochastic processes [4] with the help of kernels. This makes copula processes an appealing replacement for GPs in cases where the Gaussian assumption is not appropriate. Copula processes are relatively new in machine learning [4]. In geostatistics the copula process is called *copula based Kriging estimator* [5] and had been introduced as a possible improvement over Gaussian random fields. Multi-task learning is a more general form of co-Kriging where predictions for multiple quantities are made at the same time. Several different methods had been proposed for multi-task Gaussian processes: The task dependence can be introduced with shared hyper parameters or an appropriate prior on the covariance matrix as, for example an inverse-Wishart distribution [10]. The *Bayesian committee machine* (BCM) [8] is a *local* approximation for general probabilistic learning algorithms and belongs to the family of transductive algorithms because the predictive distribution depends on the number and location of the query points. The novelty of this work lies in the derivation of a transductive approximation for Bayesian multi-task problems.

1.1 Multi-Task Copula Processes

In contrast to single-task learning, where the objective is to estimate a scalar valued quantity, the aim of multi-task learning is to estimate more than one variable at a time.

For a finite set of input locations $X = (x_1, \ldots, x_n)$ and corresponding outputs $y = (y_1, \ldots, y_n)$ a (zero mean) Gaussian copula process $\{Y_x\}$ with marginal distribution function $F_1, \ldots, F_n$ is given as

$$p(Y_X) = c_{0,k(X,X)}(F_1(y_1), \ldots, F_n(y_n)) \cdot \prod_{i=1}^{n} \frac{\partial F_i(y_i)}{\partial y_i}, \quad (1)$$

where $c_{\mu,\Gamma}(\cdot)$ is the Gaussian copula density with mean μ and covariance Γ and $k(X, X)$ is a positive definite kernel function. In order to make predictions at input X^* we use the quantiles of the conditional distribution

$$p(Y_{X^*}|Y_X) = c_{\hat{\mu},\hat{\Gamma}}(F_1^*(y_1^*), \ldots, F_m^*(y_m^*)) \cdot \prod_{i=1}^{m} \frac{\partial F_i^*(y_i^*)}{\partial y_i^*} \quad (2)$$

$$\hat{\mu} = K(X, X^*)^T K(X, X)^{-1} w$$
$$\hat{\Gamma} = K(X^*, X^*) - K(X, X^*)^T K(X, X)^{-1} K(X, X^*)$$

and $w_i = \Phi_{0,\gamma}^{-1}(F_i(y_i))$ and Gaussian cdf Φ.

The challenge to extend a kernel-based algorithm (such as GPs or the Gaussian copula process) to a multi-task version gets reduced to the problem of defining an appropriate multi-task kernel. Some multi-task kernels are inspired from co-Kriging theory [9] as, for example the *intrinsic correlation model* (ICM) and *linear model of corregionalization* (LMC). Others are more recent such as the convolutional kernel [3].

2 Transductive Multi-Task Learning

As mentioned in the previous section, many learning algorithms, such as the ones we used in this work, can only handle a limited number training data efficiently. This makes it even harder to apply to multi-task problems, since each task carries additional data. In Kriging, Gaussian processes and Gaussian copula processes we have to do a covariance (kernel) matrix inversion, which scales cubic with the number of training data. In this section we present a transductive approach for multi-task algorithms inspired by the Bayesian committee machine [8].

Informally speaking, we are going to perform multi-task learning with the primary variable of interest and each of the secondary variables individually and combine the results at the end. This will reduce the computational costs to $\mathcal{O}(t\bar{n}^3)$.

Theorem 1. *Let $Y_{X_1}, \ldots, Y_{X_t}$ be the random variables modeling each of the t tasks and we assume without the loss of generality that we want to make predictions for the primary variable $Y_{X_1^*}$ for task 1. Using the assumption that any two Y_{X_i}, Y_{X_j} with $i \neq j \in \{2 \ldots, t\}$*

[1] A full length version of the paper is available at http://www-personal.acfr.usyd.edu.au/f.ramos/Fabio_Ramos_Homepage/Publications.html

[2] Ravensburg-Weingarten University of Applied Sciences, formerly affiliated with Australian Centre for Field Robotics
email: m.schneider@acfr.usyd.edu.au
[3] Australian Centre for Field Robotics, School of Information Technologies, The University of Sydney, Australia,
email: f.ramos@acfr.usyd.edu.au

are conditionally independent given Y_{X_1} and $Y_{X_1^*}$, we can approximate the full multi-task model as

$$P(Y_{X_1^*}|Y_{X_1},\dots,Y_{X_t}) \approxeq \frac{\prod_{i=2}^{t} P(Y_{X_1^*}|Y_{X_1},Y_{X_i})}{P(Y_{X_1^*}|Y_{X_1})^{t-2}} \cdot const.$$

Proof. See [7] for the proof □

Notice, that with this approximation, we never have to learn a model for more than two tasks at a time, which gives the computational speedup and also provides a way to easily distribute the computation to several machines.

If we apply the approximation to Gaussian copula processes, the numerator and denominator are conditional Gaussian copula densities of the form as in Eq. (2). This is advantageous since we only have to deal with products and quotients of Gaussian distributions for which analytical solutions are available. More precisely, the approximate predictive distribution for the Gaussian copula process is then

$$P(Y_{X_1^*}|Y_{X_1},\dots,Y_{X_t}) \approxeq c_{\dot{\mu},\dot{\Gamma}}(F_1^*(y_1^*),\dots,F_m^*(y_m^*)) \cdot \prod_{i=1}^{m} \frac{\partial F_i^*(y_i^*)}{\partial y_i^*},$$

where $\dot{\mu}$ and $\dot{\Gamma}$ can be obtained from

$$\mathcal{N}_{\dot{\mu},\dot{\Gamma}} = \prod_{t=2}^{t} \frac{\mathcal{N}_{\hat{\mu}_{1,i},\hat{\Gamma}_{1,i}}}{\mathcal{N}_{\hat{\mu}_{1,i},\mathrm{diag}(\hat{\Gamma}_{1,i})}} \left(\frac{\mathcal{N}_{\hat{\mu}_1,\mathrm{diag}(\hat{\Gamma}_1)}}{\mathcal{N}_{\hat{\mu}_1,\hat{\Gamma}_1}}\right)^{t-2}, \qquad (3)$$

and $\hat{\mu}_1, \hat{\Gamma}_1$, $\hat{\mu}_{1,i}, \hat{\Gamma}_{1,i}$ are defined as in Eq. (2) if we calculate the predictive distribution for $P(Y_{X_1^*}|Y_{X_1})$ and $P(Y_{X_1^*}|Y_{X_1},Y_{X_i})$ respectively. For example $\hat{\Gamma}_{1,i}$ would be obtained as

$$\hat{\Gamma}_{1,i} = K(X_1^*,X_1^*) - K([X_1,X_i],X_1^*)^T \cdot K([X_1,X_i],[X_1,X_i])^{-1} K([X_1,X_i],X_1^*),$$

which is also the main contributor to the complexity of $\mathcal{O}(8(t-1)\bar{n}^3) = \mathcal{O}(t\bar{n}^3)$. Eq. 3 above can be further reduced with the rules for products and quotients of Gaussian distributions which can be found in standard textbooks and in [8], but we omit it here due to paucity of space. Please note also that all $y_1^*,\dots,y_m^*$ are from the primary task and so are their univariate marginal distributions $F_1^*,\dots,F_m^*$.

If we follow [2], we can also see our transductive approximation as an inducing approach, where the so called *inducing variables* are defined to be Y_{X_1} and $Y_{X_1^*}$. Using this point of view, it may be easier to see that the quality of prediction can depend on the number of query points $Y_{X_1^*}$ used. As in general for transductive algorithms, the prediction becomes better, the more query points are used. As a consequence, even if only a few estimations are needed, one should include artificial dummy test inputs in the prediction step and then discard them. In most cases this is not a serious problem, since the training/parameter estimation phase is the one, which takes an order of magnitude more time than the prediction phase.

3 Experiments

The experiment is performed on the Jura dataset which contains 359 samples of two categorical variables (land uses and rock type) and the concentration of seven chemical elements. The primary variable has fewer samples than the secondary variables. This can occur in real datasets if, for example, the concentration of one element is harder or more expensive to estimate or the dataset contains missing values. For comparison reasons we use exactly the same setup as in [1].We are using the Matérn kernel for Cd, Ni and Cu and the squared exponential kernel for Zn and Pb. We are modeling the marginal distribution functions for Cd, Ni and Cu with a generalized extreme value distribution and for Zn and Pb a Gamma distribution is used.

	Opt. Time	Time/Eval.
MtGCP Cd [Ni, Zn]	898 s	0.517 s
TransGCP Cd [Ni, Zn]	429 s	0.363 s
MtGCP Cu [Pb,Ni,Zn]	1046 s	0.625 s
TransGCP Cu [Pb,Ni,Zn]	621 s	0.409 s
D200 Cd [Ni, Zn]	185 s	-
F359 Cd [Ni, Zn]	691 s	-
P200 Cd [Ni, Zn]	385 s	-

Table 1. The table shows the comparison between the full multi-task copula process (MtGCP) and the transductive approximation (TransGCP) for Cadmium (Cd) and Copper (Cu). The first column indicates the algorithm followed by the primary variable and the secondary variables in brackets. The second column shows the total time needed for the marginal likelihood optimization (Opt. Time) and the last column show the time needed per marginal likelihood function evaluation (Time/Eval). The last three entries are from [1] and the algorithm did not run on the same machine as our results. We just provide the figures for completeness and a rough baseline.

ACKNOWLEDGEMENTS

This work was supported by the Rio Tinto Centre for Mine Automation and the Australian Centre for Field Robotics.

REFERENCES

[1] M. A. Alvarez and N. D. Lawrence, 'Computationally efficient convolved multiple output gaussian processes', *Journal of Machine Learning Research*, **12**, 1425–1466, (2011).
[2] J. Q. Candela and C. E. Rasmussen, 'A unifying view of sparse approximate gaussian process regression', *Journal of Machine Learning Research*, **6**, 1939–1959, (2005).
[3] D. Higdon, 'A process-convolution approach to modelling temperatures in the north atlantic ocean', *Environmental and Ecological Statistics*, **5**(2), 173–190, (1998).
[4] S. Jaimungal and E. K. H. Ng, 'Kernel-based copula processes', *Machine Learning and Knowledge Discovery in Databases*, 628–643, (2009).
[5] H. Kazianka and J. Pilz, 'Copula-based geostatistical modeling of continuous and discrete data including covariates', *Stochastic Environmental Research and Risk Assessment*, **24**(5), 661–673, (2010).
[6] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
[7] M. Schneider and F. Ramos, 'Transductive learning for multi-task copula processes', Technical report, The University of Sydney, (2014).
[8] V. Tresp, 'A bayesian committee machine', *Neural Computation*, **12**(11), 2719–2741, (2000).
[9] H. Wackernagel, *Multivariate geostatistics: an introduction with applications*, Springer-Verlag, 2nd edn., 2003.
[10] K. Yu, V. Tresp, and A. Schwaighofer, 'Learning gaussian processes from multiple tasks', in *Proceedings of the 22nd international conference on Machine learning*, (2005).