Joint Predictive Modeling for Geospatial Data at Various Locations

Xi Cheng[1] **and Harry Xie**[2]

[1]**Department of Geography, University at Buffalo**
[2]**Department of Computer Science, University of California, Los Angeles**

Corresponding author:
Xi Cheng[1]

Email address: xcheng5@buffalo.edu

ABSTRACT

Predictive modeling uses statistics to predict unknown outcomes. In general, there are two categories of predictive modeling, parametric and non-parametric. There are many applications of predictive modeling, for example, it can be used to predict the risk score of a credit card transaction, it can also be used in health care to identify the probability of having certain disease. When it comes to geospatial data, there are some unique characteristics of the problem. Predictive modeling of geospatial data naturally involves multiple response variables at various locations. The response variables are not independent with each other and thus building separate models for each individual response variable is not appropriate. In addition, many geospatial data has strong spatial auto-correlation such that data from nearby locations are more similar with each other. A joint modeling takes into account of both the correlation among response variables and relationship among different locations, and can make predictions for locations with no training data. In this paper, we review works on joint predictive modeling for multiple response variables at various locations.

INTRODUCTION

Geospatial analysis enable us to understand the large volume of complex data and improve the decision making and plannings. For example, researchers uses constrained spectral clustering techniques Yuan et al. (2015) on a multi-scaled geospatial and temporal database Soranno et al. (2017, 2015) to make region delineations. This provides the spatial zones used in many disciplines,including economics, landscape ecology and environmental science Cheruvelil et al. (2017); Yuan et al. (2019). The analysis can also help us understand the patterns and complex relationship among geospatial ecosystems.

Nowadays, with the development of high-resolution sensing technology, geospatial data grows exponentially. Such data usually collected at multiple locations with all kinds of features. A geospatial prediction usually models different responses at different locations. For example, in Figure 1 climate scientists have data that are sampled at different location, the features can be landscape features, population information. They are interested in predicting temperature and precipitation for a future time at multiple sites simultaneously. Similarly, lake ecology researchers are interested in modeling different lake nutrients for a set of lakes Collins et al. (2019).

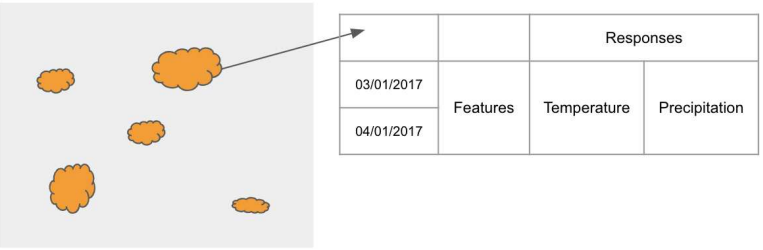

Figure 1. Joint modeling of multiple response variables at various locations.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27795v1 | CC BY 4.0 Open Access | rec: 11 Jun 2019, publ: 11 Jun 2019

One way to make the prediction is to fit one local model for each response variable at each location independent. This strategy is straight forward but suffers the following limitations:

- When the data is insufficient for each location, it is hard to build a robust model;
- This approach does not incorporate the spatial correlation between different local models;
- This model does not consider the correlations between different response variables.

To overcome these limitations, a joint predictive modeling is needed such that one can jointly make predictions for multiple responses at different locations. In this paper we review the existing works on jointly modeling of geospatial data.

1 JOINT PREDICTIVE MODELING

There are many works in developing joint predictive modeling methods Król et al. (2017); Zhao and Tang (2017); Liu et al. (2018); Yuan et al. (2017a). In Lei et al. (2015) the author proposed a method for joint learning of multiple longitudinal models for various clinical scores at multiple future time points. For every longitudinal prediction, the author adopted three relationships among training data, features, and clinical scores. The author also introduced additional relation among different longitudinal prediction models so as to select a common set of features from the baseline imaging and clinical data. The author demonstrate the effectiveness of the predictive models on Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

Many researches uses Multi-task learning (MTL) Caruana (1997) for joint predictive modeling. Multi-task learning is a class of machine learning algorithms with the advantage that it can incorporate more components of the problem into a single optimization. Rather than modeling each location as a separate learning task, Multi-task learning solves multiple, related learning tasks jointly by exploiting the common structure of the problem. Over the past decade, MTL has been successfully applied to various learning problems including regressionXu et al. (2014); Zhou et al. (2011); Yuan et al. (2017b), clustering Evgeniou et al. (2005) and classificationYu et al. (2005); Xue et al. (2007). MTL has also been applied to a wide range of applications, such as disease progression prediction Zhou et al. (2011), Web image and video search Wang et al. (2009) and Web page categorization Chen et al. (2009). The simplest way to integrate different kinds of task relationship is through a regularization term. This regularized MTL has been widely used and become a rich family of MTL. There are many MTL algorithms proposed in the literature. These algorithms vary in terms of how the task relatedness are defined and incorporated into their formulation. For example, one could assume that the model parameters for closely related tasks should be similar to each other. Such an assumption has led to the the development of the mean regularized MTL approach by Evgeniou and Pontil Evgeniou and Pontil (2004). Another common assumption is that the model parameters share a common low-rank representation. Since minimizing the rank function is an NP-hard problem, a standard approach is to minimize the trace norm instead of the rank of the model parameter matrix directly.For example, Chen et al. Chen et al. (2011) proposed a robust MTL algorithm that can identify irrelevant tasks by imposing low-rank and group-sparse constraints on the model parameters. Argyriou et al. Argyriou et al. (2008) presented a method to learn a sparse representation shared by the models for different tasks. Kumar et al.Kumar and Daume III (2012) assumed that each local model is a linear combination of finite base models.

More recently, there has been considerable interest in applying MTL to spatial, temporal, and spatio-temporal prediction problems as many of these problems can be naturally cast into a multi-task learning formulation Gonçalves et al. (2015); Xu et al. (2014, 2016a,b). For example, Xu et al. Xu et al. (2014) presented an MTL framework for ensemble time series forecasting problems. The application of the MTL framework to spatio-temporal data has also been studied by Xu et al. in Xu et al. (2016a,b). However, none of these approaches consider the nested structure of the spatial data. Another related work is the multi-level lasso approach for multi-task regression proposed by Lozano et al.Lozano and Swirszcz (2012). However, the approach assumes that the second-level variables are unobserved, unlike the formulation presented in this study, which assumes that the values of the second-level (regional) variables are observed. Since the method is not directly applicable to our geospatial predictive modeling problem, we use a variant of this multi-level lasso formulation, assuming the second-level variables are observed, as one of the baseline methods for comparing our proposed framework.

2 CONCLUSION

Geospatial data analysis is of great importance as it helps us to understand the large volume of real world data. It is common that there exist multiple correlated variables that we are intersted and thus in this paper, we reviewed techniques of joint predictive modeling for geospatial data. Specifically, we surveyed the multi-task learning framework applied in the domain of geospatial temporal database. Multi-task learning has been applied successfully across different areas, from natural language processing and computer vision to health and illness diagnosis. In this paper we reviewed many multi-tasking learning techniques on geospatial analysis. In the future, we expect more researches of other types growing in this area.

REFERENCES

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Mach. Learn.*, 73(3):243–272.

Caruana, R. (1997). Multitask learning.

Chen, J., Tang, L., Liu, J., and Ye, J. (2009). A convex formulation for learning shared structures from multiple tasks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 137–144.

Chen, J., Zhou, J., and Ye, J. (2011). Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 42–50.

Cheruvelil, K. S., Yuan, S., Webster, K. E., Tan, P.-N., Lapierre, J.-F., Collins, S. M., Fergus, C. E., Scott, C. E., Henry, E. N., Soranno, P. A., et al. (2017). Creating multithemed ecological regions for macroscale ecology: Testing a flexible, repeatable, and accessible clustering method. *Ecology and evolution*, 7(9):3046–3058.

Collins, S., Yuan, S., Tan, P.-N., Oliver, S., Lapierre, J., Cheruvelil, K., Fergus, C., Skaff, N., Stachelek, J., Wagner, T., and Soranno, P. (2019). Winter precipitation and summer temperature predict lake water quality at macroscales. *Water Resources Research*.

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.*, 6:615–637.

Evgeniou, T. and Pontil, M. (2004). Regularized multi–task learning. In *Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117.

Gonçalves, A. R., Zuben, F. J. V., and Banerjee, A. (2015). A multitask learning view on the earth system model ensemble. *Computing in Science and Engineering*, 17(6):35–42.

Król, A., Mauguen, A., Mazroui, Y., Laurent, A., Michiels, S., and Rondeau, V. (2017). Tutorial in joint modeling and prediction: A statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *Journal of Statistical Software, Articles*, 81(3):1–52.

Kumar, A. and Daume III, H. (2012). Learning task grouping and overlap in multi-task learning. In *ICML*, pages 1383–1390.

Lei, B., Chen, S., Ni, D., and Wang, T. (2015). Joint learning of multiple longitudinal prediction models by exploring internal relations. In Zhou, L., Wang, L., Wang, Q., and Shi, Y., editors, *Machine Learning in Medical Imaging*, pages 330–337, Cham. Springer International Publishing.

Liu, B., Tan, P.-N., and Zhou, J. (2018). Enhancing predictive modeling of nested spatial data through group-level feature disaggregation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1784–1793, New York, NY, USA. ACM.

Lozano, A. C. and Swirszcz, G. (2012). Multi-level lasso for sparse multi-task regression. In *Proc of Int'l Conf on Machine Learning*.

Soranno, P. A., Bacon, L. C., Beauchene, M., Bednar, K. E., Bissell, E. G., Boudreau, C. K., Boyer, M. G., Bremigan, M. T., Carpenter, S. R., Carr, J. W., et al. (2017). Lagos-ne: a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of us lakes. *GigaScience*, 6(12):gix101.

Soranno, P. A. et al. (2015). Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science through data reuse. *Giga Science*.

Wang, X., Zhang, C., and Zhang, Z. (2009). Boosted multi-task learning for face verification with applications to web image and video search. In *CVPR*, pages 142–149.

Xu, J., Tan, P., and Luo, L. (2014). ORION: online regularized multi-task regression and its application to ensemble forecasting. In *Proc of the IEEE International Conference on Data Mining*, pages 1061–1066.

Xu, J., Tan, P., Luo, L., and Zhou, J. (2016a). Gspartan: a geospatio-temporal multi-task learning framework for multi-location prediction. In *Proc of SIAM International Conference on Data Mining*, pages 657–665.

Xu, J., Zhou, J., Tan, P., Liu, X., and Luo, L. (2016b). WISDOM: weighted incremental spatio-temporal multi-task learning via tensor decomposition. In *Proc of the IEEE International Conference on Big Data*.

Xue, Y., Liao, X., Carin, L., and Krishnapuram, B. (2007). Multi-task learning for classification with dirichlet process priors. *J. Mach. Learn. Res.*, 8:35–63.

Yu, K., Tresp, V., and Schwaighofer, A. (2005). Learning gaussian processes from multiple tasks. In *Proceedings of the 22Nd International Conference on Machine Learning*, pages 1012–1019.

Yuan, S., Tan, P., Cheruvelil, K. S., Fergus, C. E., Skaff, N. K., and Soranno, P. A. (2017a). Hash-based feature learning for incomplete continuous-valued data. In *Proceedings of the 2017 SIAM International Conference on Data Mining, Houston, Texas, USA, April 27-29, 2017.*, pages 678–686.

Yuan, S., Tan, P.-N., Cheruvelil, K. S., Collins, S. M., and Soranno, P. A. (2015). Constrained spectral clustering for regionalization: Exploring the trade-off between spatial contiguity and landscape homogeneity. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Yuan, S., Tan, P.-N., Cheruvelil, K. S., Collins, S. M., and Soranno, P. A. (2019). Spatially constrained spectral clustering algorithms for region delineation. *arXiv preprint arXiv:1905.08451*.

Yuan, S., Zhou, J., Tan, P., Fergus, C. E., Wagner, T., and Soranno, P. A. (2017b). Multi-level multi-task learning for modeling cross-scale interactions in nested geospatial data. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 1153–1158.

Zhao, X. and Tang, J. (2017). Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 497–506, New York, NY, USA. ACM.

Zhou, J., Yuan, L., Liu, J., and Ye, J. (2011). A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 814–822.