Considering Autocorrelation in Predictive Models

Daniela Stojanova
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39 Ljubljana Slovenia
E-mail: Daniela.Stojanova@ijs.si, http://kt.ijs.si/daniela_stojanova

Thesis Summary

Keywords: autocorrelation, predictive models, predictive clustering trees

Received: February 21, 2013

This article presents a summary of the doctoral dissertation of the author, which addresses the task of considering autocorrelation in predictive models.

Povzetek: Članek predstavlja povzetek doktorske disertacije avtorice, ki obravnava nalogo upoštevanja autokorelacije v napovednih modelih.

1 Introduction

Most machine learning, data mining and statistical methods rely on the assumption that the analyzed data points are independent and identically distributed (i.i.d.). More specifically, the individual examples included in the training data are assumed to be drawn independently from each other from the same probability distribution. However, cases where this assumption is violated can be easily found: For example, species are distributed non-randomly across a wide range of spatial scales. The i.i.d. assumption is often violated because of the phenomenon of autocorrelation.

The cross-correlation of an attribute with itself is typically referred to as autocorrelation: This is the most general definition found in the literature. Specifically, in spatial analysis, spatial autocorrelation has been defined as the correlation among data values, which is strictly due to the relative location proximity of the objects that the data refer to. It is justified by Tobler's first law of geography [1] according to which "everything is related to everything else, but near things are more related than distant things". In network studies, autocorrelation is defined by the homophily principle [2] as the tendency of nodes with similar values to be linked with each other.

2 Methods and evaluation

In this thesis, we first give a general definition of the autocorrelation phenomenon, which includes spatial and network autocorrelation for continuous and discrete responses. We then present a broad overview of the existing autocorrelation measures for the different types of autocorrelation and data analysis methods that consider them. Focusing on spatial and network autocorrelation, we propose three algorithms that handle non-stationary autocorrelation within the tasks of classification, regression and structured output prediction. The algorithms are based on the concept of predictive clustering trees (PCTs) [3], according to which hierarchies of clusters of similar data are identified and a predictive model is associated to each cluster.

The first algorithm, SCLUS (Spatial Predictive Clustering System) [4] , explicitly considers spatial autocorrelation when learning predictive clustering models. The method is able to learn predictive models for both a continuous response (regression task) and a discrete response (classification task). It can deal with autocorrelation in the data and provide a multi-level insight into the spatial autocorrelation phenomenon. The predictive models adapt to the local properties of the data, providing at the same time spatially smoothed predictions. We evaluate SCLUS on several real world problems of spatial regression and spatial classification and show that it performs better than CLUS, which completely ignores the spatial context and CLUS*, which takes the spatial coordinates of the data points into account, but not their autocorrelation.

The second algorithm, NCLUS (Network Predictive Clustering System) [5] , explicitly considers autocorrelation when building single and multi-target predictive models from network data. We evaluate our approach on several real world problems of network regression, coming from the areas of social and spatial networks. Empirical results show that our algorithm performs better than a variety of other mainstream existing regression approaches (SVR, k-NN and M5').

The third algorithm, NHMC (Network Hierarchical Multi-label Classification) [6] , has been motivated by recent algorithms for gene function prediction where instances may belong to multiple classes and the classes are organized into a hierarchy. Thus, NHMC builds hierarchial multi-label classification models, but besides relationships among classes, it also considers the relationships among examples. Although such relationships have been identified and extensively studied in the literature, in particular as defined by protein-to-protein interaction (PPI) networks, they have not received much attention in hierarchical and

multi-class gene function prediction. The results of the evaluation show that our method, which uses the hierarchical structure of classes gene/protein properties and network information, yields better performance than the methods, using each of these sources separately.

3 Conclusion

The research presented in this thesis extends the PCT framework towards learning in the presence of autocorrelation. We consider four different types of autocorrelation (spatial, temporal, spatio-temporal and network (relational)) in the development of the proposed algorithms we focus on spatial and network autocorrelation. We address the problem of learning predictive models in the case when the examples in the data are not i.i.d, such as the definition of autocorrelation measures for a variety of learning tasks, the definition of autocorrelation-based heuristics, the development of algorithms that use such heuristics for learning predictive models, as well as their experimental evaluation on real-world data.

The major contributions of this dissertation are three extensions (SCLUS, NCLUS and NHMC) of the predictive clustering approach for handling non-stationary (spatial and network) autocorrelation for different predictive modeling tasks. The algorithms are heuristic: we define new heuristic functions that take into account both the variance of the target variables and its spatial/network autocorrelation. Different combinations of these two components enable us to investigate their influence in the heuristic function and on the final predictions. We have performed extensive empirical evaluation of the newly developed methods on single target classification and regression problems, on multi-target classification and regression problems as well as tasks of hierarchial multi-label classification.

Our approaches compare very well to mainstream methods that do not consider autocorrelation, as well as to well-known methods that consider autocorrelation. Furthermore, our approaches can more successfully remove the autocorrelation of the errors of the obtained models. Finally, the obtained predictions are more coherent in space (or in the network context). We also apply the proposed predictive models to real-word problems, such as the prediction of outcrossing rates from genetically modified crops to conventional crops in ecology, prediction of the number of views of online lectures, and protein function prediction in functional genomics.

References

[1] W. Tobler (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46:234–240.

[2] J. Neville, O. Simsek, D. Jensen (2004) Autocorrelation and relational learning: Challenges and opportunities. *Proc. Wshp. Statistical Relational Learning*, ICML, Montreal, 2004, 33–50.

[3] H. Blockeel (1998) *Top-down induction of first order logical decision trees*, PhD Thesis, Katholieke Universiteit Leuven, Belgium.

[4] D. Stojanova, M. Ceci, A. Appice, D. Malerba, S. Džeroski (2013) Dealing with spatial autocorrelation in gene flow modeling, *Ecological informatics*, 13:22–39.

[5] D. Stojanova, M. Ceci, A. Appice, S. Džeroski (2012) Network regression with predictive clustering trees, *Data Mining and Knowledge Discovery*, 25(2):378–413.

[6] D. Stojanova, M. Ceci, A. Appice, D. Malerba, S. Džeroski (2012) Using PPI networks in hierarchical multi-label classification trees for gene function prediction, *Proc. 6th International Workshop on Machine Learning in Systems Biology*, Basel, Switzerland, 10.