Learning Relations: Pitfalls and Applications

Michiel Stock, Willem Waegeman, Bernard De Baets MICHIEL.STOCK@UGENT.BE
KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics,
Ghent University, Coupure links 653, 9000 Ghent, Belgium

Keywords: learning relations, kernel methods, model evaluation, bioinformatics

Modeling the interactions between different types of agents is of importance in many domains. Social network sites are interested to predict whether two people know each other and e-commerce websites spend huge budgets to be able to recommend the right items to visitors. Also, in biology predicting interactions between different biomolecules is one of most basic steps in a systems biology approach. At a more abstract level, these different settings are identical. There are two sets of complex objects, with some kind of feature description. In addition, a relation matrix with at least some observed elements is available, denoting the interaction or relation between these objects. The problem boils down to constructing a model that is able to fill the missing values of this relation matrix.

Our methods consist of a general framework for solving these types of problems based on a joint feature representation of pairs of objects by means of the Kronecker product pairwise kernel (KPPK). In this most general case the Kronecker is taken between the two object kernels. As such a pairwise kernel matrix is obtained which encodes the covariance between two pairs of objects. This pairwise kernel matrix can subsequently be used in any kernel-based learning algorithm, such as support vector machines or kernel ridge regression. By using kernels it is possible to work with complex structured objects, such as sequences, graphs or trees. Furthermore was shown that the KPPK can be used to model arbitrary relations and can be modified for specifically learning symmetric or reciprocal relations (Waegeman et al., 2012).

Despite the strong theoretical foundations of relational learning, many open questions still remain. Currently, we are interested how to perform testing and cross validation in such settings. For example, one expects the model to make more accurate predictions for new combinations of objects that were included in the training set than for a pair of previously unseen objects. Recently in Nature, this effect was observed in a large-scale experiment (Park & Marcotte, 2012). How models for these different cases should be trained and how their performance is affected is still unclear. Furthermore we are also interested to which degree our models directly use the features of the objects, rather than indirectly exploiting the structure of the relation matrix.

A strong foundation of our framework would allow us to draw links to many other machine learning settings. Per definition, our framework can be used for general collaborative filtering problems, such as information retrieval and recommender systems. For this we have used a KPPK in an efficient ranking setting (Pahikkala et al., 2012). It is also possible to view at our models as multivariate regression or structured output prediction. If one of the types of objects could represent a certain task, our framework could be used for multitask learning. If the relation matrix is subjected to certain restrictions, it could be used for graph matching. Trying to find a generalized foundation between these different settings is an exciting part of our study.

Acknowledgments

We acknowledge the support of Ghent University (MRP Bioinformatics: from nucleotides to networks).

References

Pahikkala, T., Airola, A., Stock, M., De Baets, B., & Waegeman, W. (2012). Efficient regularized least-squares algorithms for conditional ranking on relational data. *Machine Learning*, *Submitted*.

Park, Y., & Marcotte, E. M. (2012). Flaws in evaluation schemes for pair-input computational predictions. *Nature Methods*, *9*, 1134–1136.

Waegeman, W., Pahikkala, T., Airola, A., Salakoski, T., Stock, M., & De Baets, B. (2012). A kernel-based framework for learning graded relations from data. *IEEE Transactions on Fuzzy Systems*, *99*, 1.