Mining Networked Data

Nicola Di Mauro and Donato Malerba
Department of Computer Science
University of Bari "Aldo Moro"
Bari, Italy
Email: {ndm,malerba}@di.uniba.it

ABSTRACT

The field of data mining is in the midst of a *relational revolution*. After many decades of focusing on independent and identically-distributed instances, there is a current interest for problems where instances are interdependent and linked together into a complex network. Noticeable examples of these networks are the World Wide Web, the biological networks, the sensor networks, as well as the movie and publications databases.

Networked data can be affected by several forms of autocorrelation [1], which challenges the application of both predictive and descriptive data mining methods. Informally, autocorrelation refers to the mutual conditioning of linked instances within some neighbourhood (lag). The accuracy of predictive models learned from networked data can be improved when autocorrelation is accommodated in the model, but this demands the development of both new, suitable learning methods and prediction strategies based on collective computation. At the same time, complex networks may hide different forms of autocorrelation, whose discovery opens the doors to new data mining tasks.

Current developments in the fields of Probabilistic Inductive Logic Programming (PILP) [2] and Statistical Relational Learning (SRL) [3] try to respond to this *relational revolution* by developing learning methods for rich collections of objects linked together in probabilistic relational networks. In order to apply statistical machine learning techniques to domains with complex relational and rich structure, many formalisms able to represent probabilistic relational knowledge have been proposed. With probabilistic logical languages such as Bayesian Logic Programs [4], Stochastic Logic Programs [5] or Markov Logic Networks [6] it is possible to represent different type of objects and the uncertain relations among them.

The Tutorial will cover the state of the art in this rapidly growing area of research. The goal is twofold. From one side, we intend to introduce the various forms of autocorrelation in networked data and to present the challenges that they pose to traditional data mining algorithms [7]. To this aim, we will abstract important issues from a number of application domains with various types of linked data. From the other side, we aim to provide the audience with a survey and a comparison of different SRL representations, distinguishing among different SRL tasks and presenting SRL applications. Special emphasis will be given to SRL algorithms for two logic based formalisms: Markov Logic Networks (MLN) and Logic Programs with Annotated Disjunctions (LPAD) [8], [9]. Finally, intersections with contributions from the Computational Intelligence side will be discussed and possible research directions will be outlined.

REFERENCES

[1] D. Jensen and J. Neville, "Linkage and autocorrelation cause feature selection bias in relational learning," in *Proceedings of the Nineteenth International Conference on Machine Learning*, C. Sammut and A. G. Hoffmann, Eds. Morgan Kaufmann, 2002, pp. 259–266.

[2] L. D. Raedt and K. Kersting, "Probabilistic inductive logic programming," in *Probabilistic Inductive Logic Programming*, ser. Lecture Notes in Computer Science, L. D. Raedt, P. Frasconi, K. Kersting, and S. Muggleton, Eds., vol. 4911. Springer, 2008, pp. 1–27.

[3] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[4] K. Kersting and L. D. Raedt, "Basic principles of learning bayesian logic programs," in *Probabilistic Inductive Logic Programming*, ser. Lecture Notes in Computer Science, L. D. Raedt, P. Frasconi, K. Kersting, and S. Muggleton, Eds., vol. 4911. Springer, 2008, pp. 189–221.

[5] S. Muggleton, "Learning structure and parameters of stochastic logic programs," in *12th International Conference on Inductive Logic Programming*, ser. Lecture Notes in Computer Science, S. Matwin and C. Sammut, Eds., vol. 2583. Springer, 2002, pp. 198–206.

[6] M. Richardson and P. Domingos, "Markov logic networks," *Machine Learning*, vol. 62, no. 1-2, pp. 107–136, 2006.

[7] A. Appice, M. Ceci, and D. Malerba, "An iterative learning algorithm for within-network regression in the transductive setting," in *Proceedings of the 12th International Conference on Discovery Science*. Springer, 2009, pp. 36–50.

[8] J. Vennekens, S. Verbaeten, and M. Bruynooghe, "Logic programs with annotated disjunctions," in *Proceedings of the 20th International Conference on Logic Programming*, ser. Lecture Notes in Computer Science, B. Demoen and V. Lifschitz, Eds., vol. 3132. Springer, 2004, pp. 431–445.

[9] F. Riguzzi and N. Di Mauro, "Applying the information bottleneck to statistical relational learning," *Machine Learning Journal*, 2011.

[10] L. D. Raedt, P. Frasconi, K. Kersting, and S. Muggleton, Eds., *Probabilistic Inductive Logic Programming - Theory and Applications*, ser. Lecture Notes in Computer Science, vol. 4911. Springer, 2008.