Statistical Relational Learning for Clinical Domains

Jesse Jon Davis

Department of Computer Science
Katholieke Universiteit Leuven

Abstract. Machine learning has become an essential tool for analyzing biological and clinical data, but significant technical hurdles prevent it from fulfilling its promise. Standard algorithms make two key assumptions: the training data consist of independent examples and each example is described by a pre-defined set of attributes. Biomedical domains consist of complex, inter-related, structured data, such as patient clinical histories, molecular structures and protein-protein interaction information. The representation chosen to store the data often does not explicitly encode all the necessary features and relations for building an accurate model. For example, when analyzing a mammogram, a radiologist records many properties of each abnormality, but does not explicitly encode how quickly a mass grows, which is a crucial indicator of malignancy. This talk will describe an approach that automatically discovers unseen features and relations from data, which has advanced the state-of-the-art for machine classification of abnormalities on a mammogram. Presently most of the women identified for a possible malignancy on a mammogram are called back unnecessarily, with concomitant stress, procedure (additional imaging and/or biopsy) and expense. This research, which achieves superior performance compared to both previous machine learning approaches and radiologists, has demonstrated the potential to dramatically reduce this fraction without reducing the number of cancers correctly diagnosed.

Overview

Statistical relational learning (SRL) [5, 6], which combines first-order logic with probability, can model the complex, uncertain, structured data that characterizes clinical and biological domains. In these types of problems, the available data often does not contain all the necessary features and relations for building an accurate model. However, most SRL algorithms are constrained to use a pre-defined set of features and relations during learning. Requiring a domain expert to hand-craft all relevant features or relations necessary for a problem is a difficult and often infeasible task. Ideally, the learning algorithm should automatically discover and incorporate relevant features and relations.

The Score As You Use (SAYU) algorithm [1–4] is a general SRL framework for discovering new features and relations during learning. SAYU defines features and relations as first-order logical rules and evaluates each one by building a new statistical model that incorporates it. If adding the new feature or relation improves the model's predictive performance, then it is retained in the model. SAYU has been successfully applied to several tasks, including diagnosing breast cancer from structured mammography reports and predicting three-dimensional Quantitative Structure-Activity Relationships (3D-QSAR) for drug design.

Labeling an abnormality as benign or malignant from a structured mammography report is a challenging task for both radiologists and machines. Previous machine learning approaches to this problem are limited to using pre-defined features and ignore the relational nature of this task. However, a radiologist might include derived features which incorporate data about other abnormalities on the same mammogram or prior abnormalities in the decision process. SAYU, which can construct additional features that incorporate relational information, significantly outperformed radiologists, a hand-crafted Bayesian network system, standard Bayesian network structure learners and other SRL systems [2, 4] for this task.

References

1. Jesse Davis, Elizabeth Burnside, Inês Dutra, David Page, and Vítor Santos Costa. An integrated approach to learning Bayesian networks of rules. In *Proceedings of the 16th European Conference on Machine Learning*, pages 84–95. Springer, 2005.
2. Jesse Davis, Elizabeth Burnside, Inês C. Dutra, David Page, Raghu Ramakrishnan, Vítor Santos Costa, and Jude Shavlik. Learning a new view of a database: With an application to mammography. In Lise Getoor and Ben Taskar, editors, *An Introduction to Statistical Relational Learning*. MIT Press, 2007.
3. Jesse Davis, Vítor Santos Costa, Soumya Ray, and David Page. An integrated approached to feature construction and model building for drug activity prediction. In *Proceedings of the 24th International Conference on Machine Learning*, pages 217–224. ACM Press, 2007.
4. Jesse Davis, Irene Ong, Jan Struyf, Elizabeth Burnside, David Page, and Vítor Santos Costa. Change of representation for statistical relational learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2719–2726. AAAI Press, 2007.
5. Luc De Raedt, Paolo Frasconi, Kristian Kersting, and Stephen Muggleton, editors. *Probabilistic inductive logic programming: theory and applications*. Springer-Verlag, Berlin, Heidelberg, 2008.
6. Lise Getoor and Ben Taskar, editors. *An Introduction to Statistical Relational Learning*. MIT Press, 2007.