Mutagenesis using Inductive Logic Programming Project Proposal

Arunothia Marappan *

*Indian Institute of Technology Kanpur

Submitted to Prof. Amitabha Mukherjee for partial fulfilment of the course requirements for CS365A, IITK

Mutagenicity refers to a chemical or physical agent's capacity to cause mutations (genetic alterations) [11].Being able to Predict the mutagenicity of a compound is very essential to prevent deadly diseases like cancer. Laboratory testing for mutagenicity is a very costly and a time consuming task and in some cases it is not even achievable as of date.Therefore, building a learning algorithm that can solve this problem using a set postive and negtaive examples as its training is the requirement. The complexity of the problem rises with the scarcity of the domain knowledge available. Inductive Logic Programming (ILP) remains one of the best approaches to this problem as induction seems to be the only understandable way to arrive at a solution when there is not much domain knowledge available with us.

Bond-level data | heuristics | prolog Implementation | progol Algorithm

Problem Statement

In this project, I aim at providing a machine learning algorithm for the classification of the chemical compounds based on their mutagenicity. The classification will be binary and will be returning true if the compound shows more chances of being mutagenic. I will use the approach of Inductive Logic Programming to achieve the same. Different approaches to ILP will be learnt and tested and the best approach will be presented.

Motivation

The main motivation is to learn and implement empirical ILP using prolog. To learn and understand how to find the best pruning strategies and heuristics from the given background knowledge.

Dataset

Bond-level data. The atom and bond structures of the 230 drugs were obtained from the standard molecular modelling package QUANTA. For each compound QUANTA automatically obtains the atoms, bonds, bond types (for example, aromatic, single, double etc.), atom types (for example, aromatic carbon, aryl carbon etc.), and the partial charges on atoms. QUANTA auto-matically classifies bonds into one of 8 types, and atoms into one of 233 types. The output was a set of Prolog facts which will serve as the database for this project.The resulting 12203 facts on atomic structure and bonding generated by QUANTA form the only knowledge available for learning. Of the 230 compounds, 138 have positive levels of log mutagenicity and the rest have negative mutagenicity. [1]

Methodology

Inductive Logic programming(ILP) is a subfield of machine learning which uses logic programming as a uniform representation for examples, background knowledge and hypotheses. Given an encoding of the known background knowledge and a set of examples represented as a logical database of facts, an ILP system will derive a hypothesised logic program which entails all the positive and none of the negative examples. [11]
[3] An ILP learner can be described by

- the structure of the space of clauses
- its search strategy
  - uninformed search (depth-first, breadth-first, iterative deepening)
  - heuristic search (best-first, hill-climbing, beam search)
- its heuristics
  - for directing search
  - for stopping search (quality criterion)

In their paper - *ILP Experiments in a non determinate biological domain* S.H.Muggleton and A.Srinivasan [1], have described an algorithm named *progol* which does exactly what I aim to do in this project. But, there are many other ways too to approach this problem. Any valid ILP model will work, but its efficiency will vary. I aim to analyse some common models and present the best of them using the programming language *prolog*.

1. A.Srinivasan and S.H Muggleton, Mutagenesis: ILP Experiments in a non determinate biological domain
2. Ashwin Srinivasan and S.H Muggleton, Theories for mutagenicity: a study in first-order and feature-based induction
3. Notes by J.Stephen, Introduction to ILP
4. (Lecture notes in computer science 4455) Stephen Muggleton, Ramon Otero, Alireza Tamaddoni-Nezhad-Inductive Logic Programming, 16 conf., ILP 2006-Springer (2007)
5. (Lecture notes in computer science Lecture notes in artificial intelligence 1228) Shan-Hwei Nienhuys-Cheng, Ronald de Wolf-Foundations of Inductive Logic Programming-Springer (1997)
6. Lecture Notes in Computer Science, 7207 Lecture Notes in Computer Science, 7207 Stephen H Muggleton Alireza Tamaddoni-Nezhad Francesca A Lisi-Inductive Logic Programming 21st International Conf
7. Frontiers in Artificial Intelligence and Applications, Vol. 148 K. Kersting-An Inductive Logic Programming Approach to Statistical Relational Learning Volume 148 Frontiers in Artificial Intelligence
8. http://www-ai.ijs.si/ ilpnet2/education/index.html
9. http://www.doc.ic.ac.uk/ shm/ilp.html
10. http://www-users.cs.york.ac.uk/ jc/teaching/GSLT/ilp-goth.html
11. Wikipedia

Reserved for Publication Footnotes