KDD CUP 2001 Task 1: Thrombin

Jie Cheng

(www.cs.ualberta.ca/~jcheng)

Global Analytics

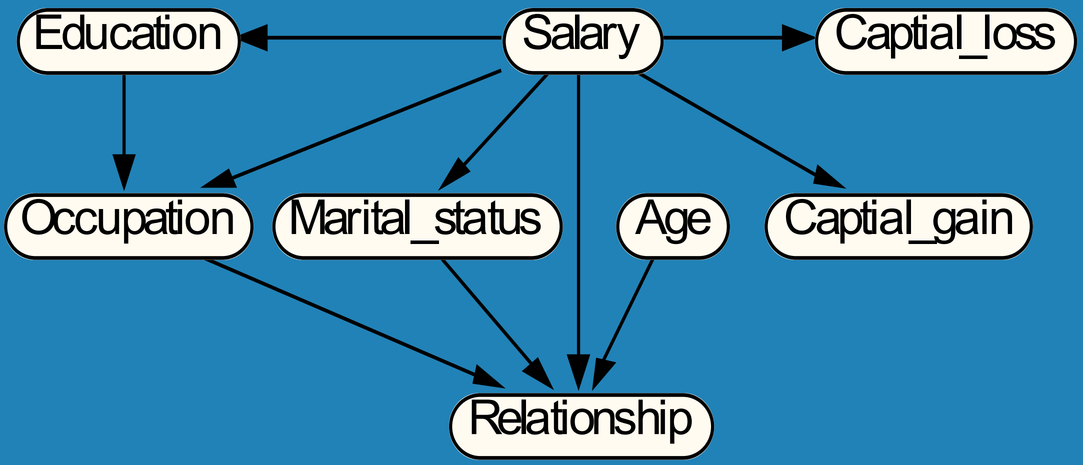
Canadian Imperial Bank of Commerce

Overview

- Objective
  - Prediction of molecular bioactivity for drug design -- binding to Thrombin
- Data
  - Training: 1909 cases (42 positive), 139,351 binary features
  - Test: 634 cases
- Challenge
  - Highly imbalanced, high-dimensional, different distribution
- My approach
  - Bayesian network predictive model

Bayesian Networks (BN)

- What is a Bayesian Network
- Two ways to view it:
  - Encodes conditional independent relationships
  - Represents the joint probability distribution



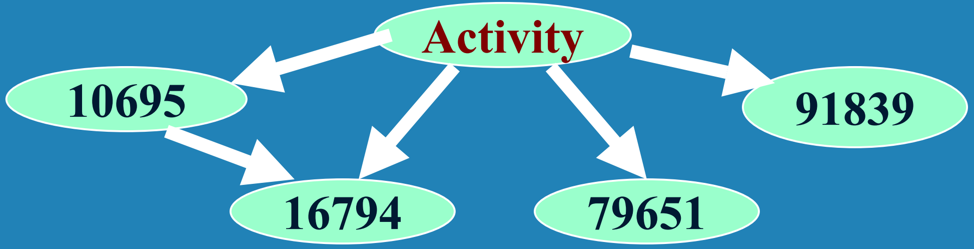
My work related to BN

- Developed an efficient approach to learn BN from data (paper to appear in Artificial Intelligence Journal)
- ***BN PowerConstructor system***
  - available since 1997, thousands of downloads and many regular users
- Learning BNs as predictive models
- ***BN PowerPredictor system***
  - available since 2000
- Applied BN learning to UCI benchmark datasets, Power plant fault diagnosis, Financial risk analysis

Approach to Thrombin data

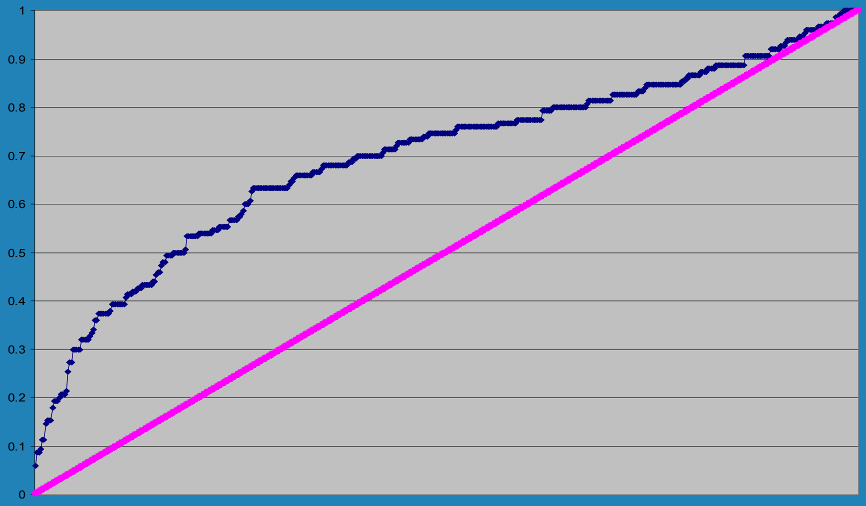
- Pre-processing: Feature subset selection using mutual information (200 of 139,351 features)
- Learning Bayesian network models of different complexity (2 to 12 features)
- Choosing a model (ROC area, model complexity)
- Cost function?
  - From posterior probability: only 10 cut points: 30, 31, 32, 71, 72, 74, 75, 215, 223, 550

The model & its performance



20 parameters

ROC



		predicted	
		pos	neg
Actual	pos	95	55
	neg	128	356

Accuracy: 0.711
Weighted Accuracy: 0.684

Bayesian networks make good classifiers!

- Accurate
  - UCI datasets
- Efficient
  - Learning: linear to number of samples, O(N^2) to the number of features – seconds to minutes
  - Inference: simple multiplications
- Markov blanket for feature selection
- Easy to understand, easy to incorporate domain knowledge

BN PowerPredictor System

- Download:

  http://www.cs.ualberta.ca/~jcheng/bnsoft.htm

- Features:
  - Support domain knowledge input
  - Support multiple database and spreadsheet formats
  - Automatic feature subset selection
  - Automatic model selection using wrapper approach
  - provide equal size, equal frequency and entropy based discretization
  - Support cost function definition
  - Instant/batch inference models