Applications of Bayesian Networks

Yuqing Tang

Doctoral Program in Computer Science
The Graduate Center
City University of New York
ytang@cs.gc.cuny.edu

November 24, 2010

Introduction

[Pourret *et al.*, 2008]

- Medical domain: e.g. Medical Diagnosis & Clinical Decision Support
- Scientific domain: e.g. Complex Genetic Models
- Crime and terrorism management domain: e.g. Risk factors analysis, Inference in Forensic Science, Terrorism Risk Management
- Social domain: e.g. Spatial Dynamics, Student Modeling
- Mining: e.g. Classifiers for Modeling of Mineral Potential
- Financial and business domain: e.g. Credit-Rating of Companies, Predicting Probability of Default for Large Corporates
- Manufacture monitor and control domain: e.g. Reliability Analysis of Systems with Dynamic Dependencies, Decision Support on Complex Industrial Process Operation
- Information retrieval domain: e.g. An Information Retrieval System for Parliamentary
- Robotics: e.g. Risk Management in Robotics
- Others: e.g. Classification of Chilean Wines

Outline

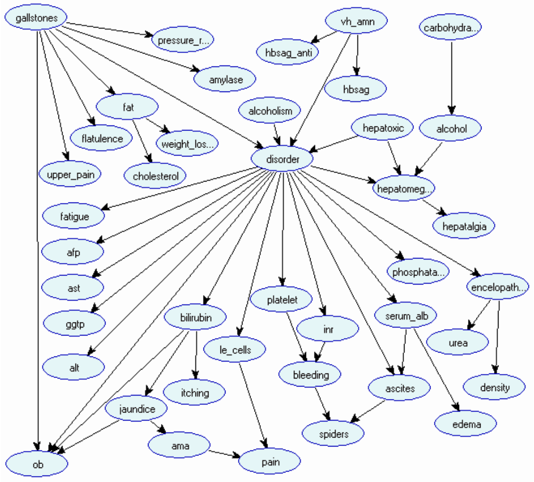
1. Applications

2. A list of readings

3. Summary

Medical Diagnosis I

A Probabilistic Causal Model for Diagnosis of Liver Disorders
[Pourret *et al.*, 2008, Chapter 2]



Medical Diagnosis II

- Problem setting
  - The starting point for building our model has been HEPARs database of patient cases
  - The database available to us included 570 patient records, each of these records was described by 119 features (binary, denoting presence or absence of a feature or continuous, expressing the value of a feature) and each record belonged to one of 16 liver disorders
  - One limitation of the HEPAR database is the assumption that a patient appearing in the clinic has at most one disorder
  - The features can be divided conceptually into three groups: symptoms and findings volunteered by the patient, objective evidence observed by the physician, and results of laboratory tests
- Attacking the problem
  - Initial attempt to reduce the number of features from the 119 encoded in the database to 40

Medical Diagnosis III

  - Then rely on experts opinion as to which features have the highest diagnostic value. Having selected the total of 40 features, the authors elicited the structure of dependences among them from our domain experts.

- Model parameters
  - Continuous variable discretization is based on expert opinion that variables such as urea, bilirubin, or blood sugar have essentially low, normal, high, and very high values. The numerical boundaries of these intervals are based on expert judgment.
  - The program learns from HEPAR database the parameters of the network, i.e., prior probabilities of all nodes without predecessors and conditional probabilities of all nodes with predecessors, conditional on these predecessors.
    - ★ Prior probability derived from the relative counts of various outcomes for each of the variables in question
    - ★ Conditional probability distributions are relative counts of various outcomes in those data records that fulfill the conditions described by every combination of the outcomes of the predecessors.

Medical Diagnosis IV

        - ★ Interpret the missing measurements as possible values of the variables in question

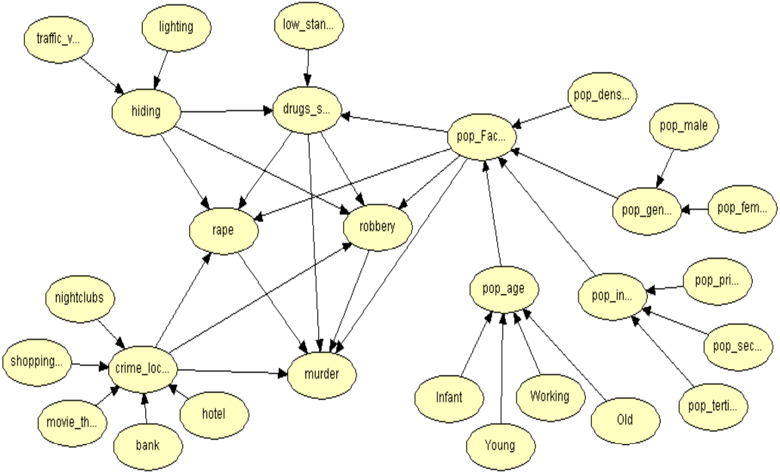
- Evaluating the performance
    - The authors used a fraction of the database to learn the network parameters and the remainder of the records to test the network prediction.
    - In over 36% of the cases, the most likely disorder indicated the correct diagnosis. In over 74% of the cases, the correct diagnosis was among the first four most likely disorders, as indicated by our model.
    - Some points: The causal model may perform worse in numerical terms than a regression-based model,1 it offers three important advantages: (1) its intuitive and meaningful graphical structure can be examined by the user, (2) the system can automatically generate explanations of its advice that will follow the model structure and will be reasonably understandable, and (3) the model can be easily enhanced with expert opinion; interactions absent from the database can be added based on knowledge of local causal interactions with the existing parts and can be parameterized by expert judgment.

Crime Risk Factors Analysis I

[Oatley and Ewart, 2003] and [Pourret *et al.*, 2008, Chapter 5] The following are taken from [Pourret *et al.*, 2008, Chapter 5]

- Crime pattern analysis can guide planners in the allocation of resources
- Requirements
  - Make accurate predictions
  - Accommodate changes in various parameters over time
- Data set
  - A set of data consisting of 1000 records
  - 20 variables which are classified in five groups: population, crime locations, types of crimes, traffic, and environment
- The BN is learned from the data and an initial structure that reflects the human experts' opinions
- Performance suggested that machine learning techniques can be used to analyze crime data and help in crime control planning

Crime Risk Factors Analysis II



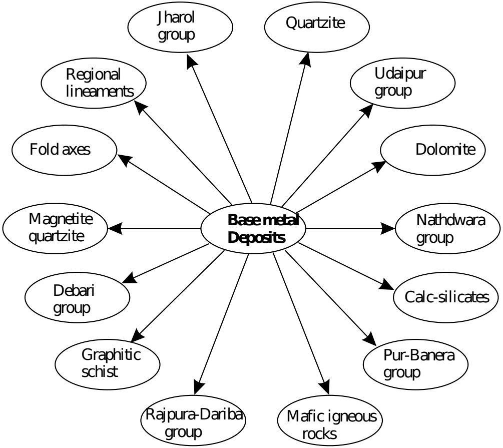
Spatial Dynamics model through Bayesian Networks

[Pourret *et al.*, 2008, Chapter 6]: Spatial Dynamics in the Coastal Region of South-Eastern France (pages 87111)

- Spatial databases + modelers' knowledge ⇒ models on the dynamics of European metropolitan areas in the last decades (as of 2008)
- Three French areas: Marseilles, Aix-en-Provence and Toulon
- The basis of the information: Globalization, competition among cities, the development of inter-metropolitan networks, and the concentration of rare functions
- The goal: The analysis of the spatial dynamics characterizing the emergence of the metropolitan systems at the local level
- Variables
  - 39 indicators

Classifiers for Modeling of Mineral Potential

[Pourret *et al.*, 2008, Chapter 9]



Graphitic schist	Deposit	Rajpura-Dariba group	
		Absent	Present
Absent	Absent	1.000	0.000
Absent	Present	0.907	0.093
Present	Absent	1.000	0.000
Present	Present	0.300	0.700

Student Modeling I

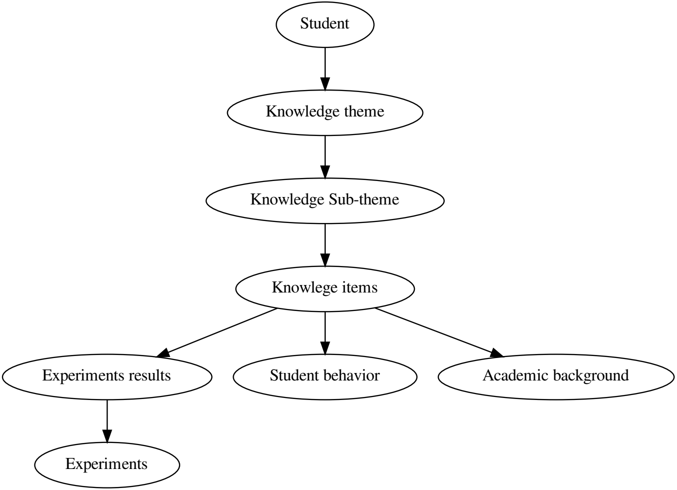
[Pourret *et al.*, 2008, Chapter 10]

- Modeling students so that intelligent tutoring systems (virtual laboratories) can adapt to the learner
- Probabilistic relational models: an extension of Bayesian networks
  - Entities are objects or domain entities that are partitioned into a set of disjoint classes $X_1, ..., X_n$
  - Each class X_i is associated with a set of attributes $A(X_i) = \{A_{i,j}\}$ which can take a fixed domain of values $V(A_{i,j})$
- School domain
  - Professor: teaching-ability, popularity
  - Student: intelligence, ranking
  - Course: rating, difficulty
  - Registration: satisfaction, grade

Student Modeling II

- The model keeps track of student's knowledge at different levels of granularity, combing the performance and exploration behavior in several experiments, to decide the best way to guide and to recategorize the student

Student Modeling III

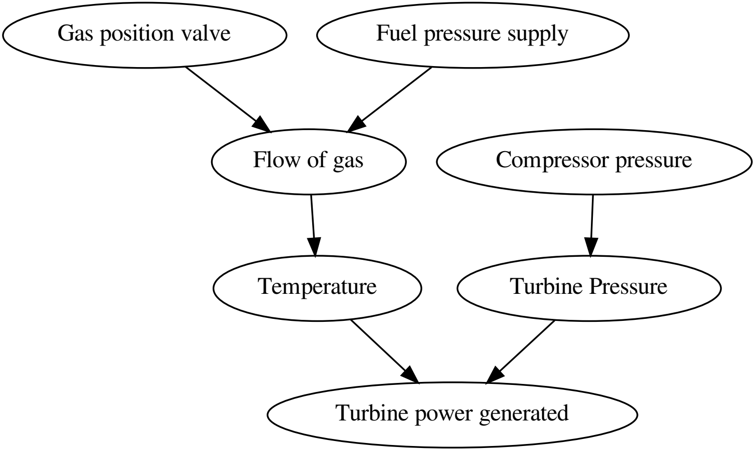


Sensor Validation I

[Pourret *et al.*, 2008, Chapter 11]

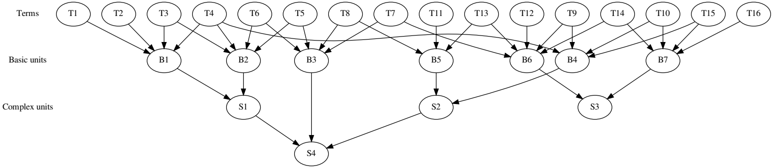
- Complex equipment and instrumentation are used to constantly monitor the status of the environment and the behaviors of a system
- Validation of sensors
  - In the first phase, potential faults are detected by comparing actual sensor value with the one predicted from the related sensors, via propagation in the Bayesian network
  - In the second phase, the real faults are isolated by constructing an additional Bayesian network based on the Markov blanket property
  - This isolation is made incrementally (any time), so the quality of the estimation increases when more time is spent in the computation (suitable for use in real-time environments)

Sensor Validation II



Information Retrieval Systems

[Pourret *et al.*, 2008, Chapter 12]



- Information retrieval
  - User expresses his/her information need using a natural language query
  - The systems output a set of relevant documents sorted according to its degree of relevance
    - One step further: output the relevant parts sorted according to its degree of relevance

Credit-Rating of Companies I

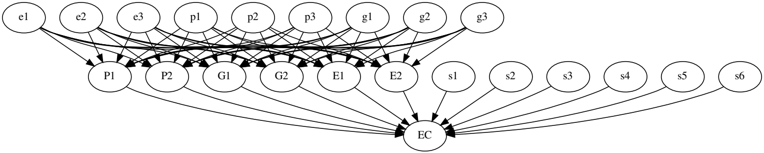
[Pourret *et al.*, 2008, Chapter 15]

- Credit-rating or appraisal of companies is the evaluation of desirability of companies as targets of investment
  - Credit-ratings are based primarily on much current financial data on performance of companies
  - Unofficial/insider information, current market conditions and past data are also involved
- Data: 523 complete cases of credit-rating data by experts and 15 official financial indexes
- Human process
  - 16 indexes are factored into 4 major factors
    - Profit performance
    - Financial security
    - Growth potential
    - Firm size
- Several Bayesian networks are constructed
  - Some are constructed following financial experts' points of view

Credit-Rating of Companies II

  - Naive Bayesian classifiers: assume that all the feature variables are conditionally independent on each other given the class variable

- Leave-one-out cross-validation test on each model



Profit performance (p), growth potential (g), corporate efficiency and productivity (e), and financial security (s). P1 and P2, G1 and G2, and E1 and E2 are the principle components of the corresponding factors. EC is the rating classes.

Outline

1. Applications

2. A list of readings

3. Summary

A list of readings I

- A Probabilistic Causal Model for Diagnosis of Liver Disorders
  - http://www.google.com/url?sa=t&source=web&cd=3&ved=0CCUQFjAC&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.79.9931%26rep%3Drep1%26type%3Dpdf&ei=w0boTMmJ0IH88Aa0sZnhDA&usg=AFQjCNGmzxy3Itk8TU3uRW9ZXkFiuKkWkQ&sig2=Qd1k0fIAG62yogFoWSnJ9w
- TakeHeart II: A tool to support clinical cardiovascular risk assessment
  - http://www.google.com/url?sa=t&source=web&cd=2&ved=0CCUQFjAB&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.69.7994%26rep%3Drep1%26type%3Dpdf&rct=j&q=decision%20support%20for%20clinical%20carniovascular%20risk%20assessment&ei=q0roTJHpAo0C8gbq7dWoCQ&usg=AFQjCNHfhBhzQ78DqZh3EoYqwWFG5BXQBg&sig2=MfqQtSFhLEVWD_tkoSUHHA

A list of readings II

- A Bayesian Network Model for Analysis of the Factors Affecting Crime Risk
  - `http://www.wseas.us/e-library/conferences/crete2004/papers/476-305.pdf`
- Conceptual modelling of the interaction between transportation, land use and the environment as a tool for selecting sustainability indicators of urban mobility
  - `http://cybergeo.revues.org/index1590.html`
- Mineral Potential Mapping with Mathematical Geological Models
  - `http://www.itc.nl/library/papers_2006/phd/porwal.pdf`
- A Probabilistic Model for Information and Sensor Validation
  - `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.109.5169`
- Bayesian networks and information retrieval: an introduction to the special issue

A list of readings III

  - http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VC8-4C6TF0P-1&_user=699479&_coverDate=09/30/2004&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000039280&_version=1&_urlVersion=0&_userid=699479&md5=09d336ec2f48dd0b7c279a9e0eb5464f&searchtype=a

- Compiling Dyanamic Fault Trees into Dynamic Bayesian Nets for Reliability Analysis: the Radyban Tool
  - http://www.google.com/url?sa=t&source=web&cd=5&ved=0CDYQFjAE&url=http%3A%2F%2Fvolgenau.gmu.edu%2F~klaskey%2Fuai07workshop%2FAppWorkshopProceedings%2FUAIAppWorkshop%2Fpaper6.pdf&ei=y1HoTKDsD4H-8Abf3_GiDQ&usg=AFQjCNFs3cUxTxjCti8_M9bii5Br4XZ8nA&sig2=RKzMGzXoYDblFTmyK_KN8g
- State/Local CIP Risk Analysis: First Results and Emerging Trends in the Data
  - http://www.dsbox.com/images/uploads/Daniels_benchmarking.pdf

A list of readings IV

- Wijayatunga, P., Mase, S. and Nakamura, M. "Appraisal of Companies with Bayesian Networks", (2006) International Journal of Business Intelligence and Data Mining, Vol. 1, No. 3, pp.326346.
  - `http://www3.umu.se/stat/personal/priyantha.wijayatunga/IJBIDM1(3)Paper4.pdf`
- Object Oriented Bayesian Networks for Industrial Process Operation
  - `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.5351&rep=rep1&type=pdf`
- The Credit Rating Process and Estimation of Transition Probabilities: A Bayesian Approach
  - `http://faculty.london.edu/cstefanescu/Stefanescu-Tunaru-Turnbull.pdf`
- Applications of Probabilistic Graphical Models to Diagnosis and Control of Autonomous Vehicles

A list of readings V

- http://www.google.com/url?sa=t&source=web&cd=8&ved=0CEcQFjAH&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.141.2667%26rep%3Drep1%26type%3Dpdf&rct=j&q=risk%20management%20in%20robotics%20madsen%20kalwa%20&ei=K1joTPGdD8H48Aaa8KmZDQ&usg=AFQjCNHus0tgS6dAPNXdigcSXftCh9xB0A&sig2=kTnNGEo_TkkxotyeyTDoJA

Outline

1. Applications

2. A list of readings

3. Summary

Summary

- Comparing with expert systems
  - Bayesian networks: Knowledge on the domain
  - Expert systems: Knowledge on the expert
- Bayesian networks
  - Acquiring knowledge from data
  - Assessing uncertainty
  - Combing knowledge sources
  - Putting knowledge to work
- Limitations
  - Computational complexity: exponential in terms of connectivity of the networks
  - Not an issue in expert based networks: most knowledge based expressed by human experts will stay within a reasonable size limit
- Challenges
  - Intelligence not limited to inference and learning
  - The requirement of actions
  - Embedded AI in real work (not just an offline AI system connecting to sensors and actuators)

Acknowledgments

The materials of the slides are taken from [Pourret *et al.*, 2008] with the instructor's own interpretation.

References I

Giles C. Oatley and Brian W. Ewart.
Crimes analysis software: "pins in maps", clustering and bayes net prediction.
Expert Systems with Applications, 25(4):569 – 588, 2003.

Olivier Pourret, Patrick Naim, and Bruce Marcot, editors.
Bayesian networks: A practical guide to applications.
John Wiley and Sons, 2008.