Multivariate Conditional Anomaly Detection and Its Clinical Application

Charmgil Hong (Academic advisor: Milos Hauskrecht) / charmgil@cs.pitt.edu / Department of Computer Science, University of Pittsburgh

Introduction

- **Motivation from the clinical domain** [James '13]
  - Preventable Medical errors are estimated to be approx. 210k-440k patients/year
  - This is the third leading causes of death in America

- **Motivation from existing computer-based decision supporting systems**

(1) Knowledge-driven approach

- A number of solutions exist; primarily built by medical/clinical experts
- These solutions are usually very expensive and their coverage is rather incomplete
- e.g. A Bayesian network for liver disorder diagnosis [Onisko et al. '99]

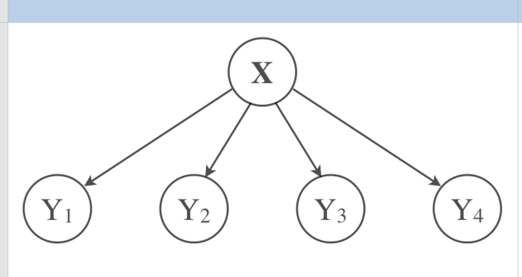
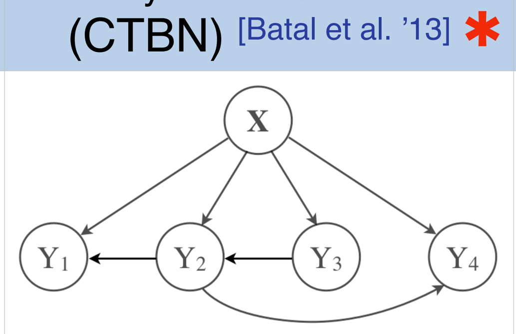
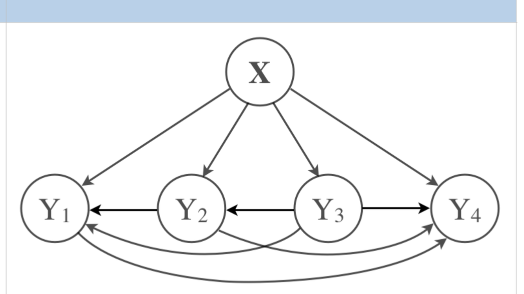
(2) *Data-driven approach**

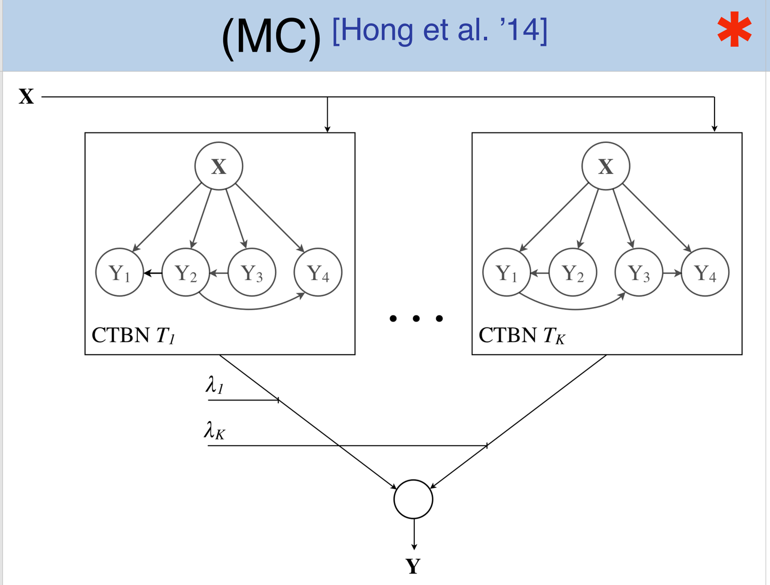
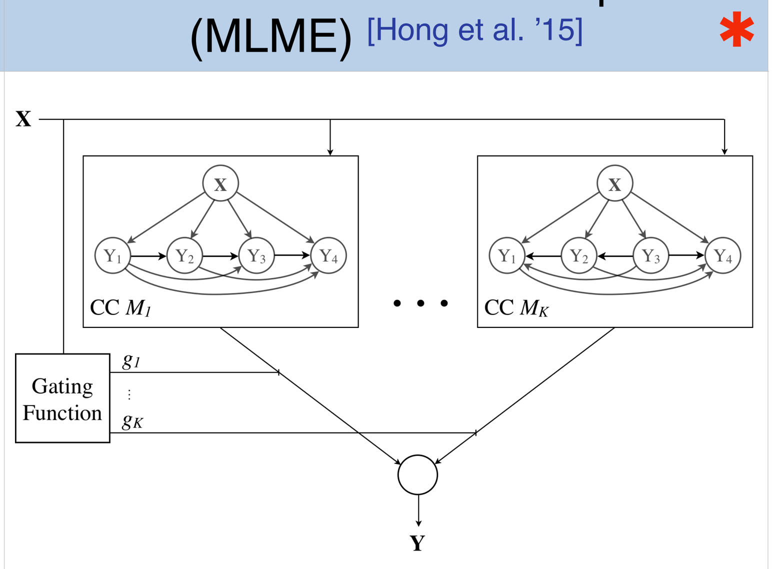
- Medical errors can be thought as statistical anomalies based on past clinical data stored in electronic medical record (EMR) systems
- Cases requiring medical attention for reconsideration could be identified by detecting anomalies in patient care patterns

Approach

Phase 1: Multi-dimensional Data Modeling of Clinical Records

- Objective:
  - Model a conditional joint distribution $P(\mathbf{y}|\mathbf{x})$ of clinical actions $\mathbf{y} = \{y_1, ..., y_d\}$ (output) given patient condition $\mathbf{x} = \{x_1, ..., x_m\}$ (input)
  - Learn a function that assigns to each patient condition $\mathbf{x}$, the most probable (MAP; maximum a posteriori) assignment of the clinical actions $\mathbf{y}$
- Challenge: The number of all possible class assignments is exponential in $d = |\mathbf{Y}|$
- Solutions (✱ *indicates our contributions*)

Model	Binary Relevance (BR) [Boutell et al. '04]	Conditional Tree-structured Bayesian Networks (CTBN) [Batal et al. '13] ✱	Classifier Chains (CC) [Read et al. '09]
Graphical Representation (e.g., $d = 4$)			
Mathematical Representation	$P(\mathbf{Y}\mid\mathbf{X}) = \prod_{i=1}^{d} P(Y_i\mid\mathbf{X}, \pi(Y_i))$		
	$\pi(Y_i) = \{\}$	$\pi(Y_i)$ = at most one parent label (tree)	$\pi(Y_i)$ = all preceding labels (chain)
Strength	· Structure learning is not required (fast)	· Optimal tree structures are learned efficiently · Exact MAP inference can be performed in a linear time (Max-sum)	· Theoretically, CC does not lose any class dependency (∵ chain rule)
Weakness	· BR disregards all the class dependencies · It is a simple collection of marginal models	· The dependency can be learned is limited to a tree structure	· Learning the optimal structure is NP-hard · Exact MAP inference is NP-hard · A greedy approx. is used

Mixture	Mixtures-of-CTBNs (MC) [Hong et al. '14] ✱	Multi-label Mixtures-of-Experts (MLME) [Hong et al. '15] ✱
Graphical Representation (e.g., $d = 4$)		
Mathematical Representation	$P(\mathbf{y}\mid\mathbf{x}) = \sum_{k=1}^{K} \lambda_k \prod_{i=1}^{d} P(y_i\mid\mathbf{x}, \pi(y_i, T_k))$	$P(\mathbf{y}\mid\mathbf{x}) = \sum_{k=1}^{K} g_k(\mathbf{x}) \prod_{i=1}^{d} P(y_i\mid\mathbf{x}, \boldsymbol{\pi}(y_i, M_k))$
	λ_k: (fixed) weight of the k-th model	$g_k(\mathbf{x})$: weight of the k-th model given $\mathbf{x}$
Strength	· Can have multiple dependency structures	· Can take any of BR, CTBN, CC as the base structures
Weakness	· Only permits CTBNs as the base structures	· Computationally more demanding · Requires to learn the gating function along with the k models

- **Why are the marginal models not enough?**

Given the joint probability table below, find the most probable assignment (MAP; maximum a posteriori) of $\mathbf{Y} = (Y_1, Y_2)$

$P(Y_1,Y_2\mid\mathbf{X}{=}\mathbf{x})$	$Y_1 = 0$	$Y_1 = 1$	$P(Y_2\mid\mathbf{X}{=}\mathbf{x})$
$Y_2 = 0$	0.2	0.45	**0.65**
$Y_2 = 1$	0.35	0	0.35
$P(Y_1\mid\mathbf{X}{=}\mathbf{x})$	**0.55**	0.45	

→ Prediction on the joint (MAP): $Y_1 = 1$, $Y_2 = 0$

→ Prediction on the marginals: $Y_1 = 0$, $Y_2 = 0$

Phase 2: Estimating Anomaly Scores

- Objective
  - Given a trained model and unseen test data, precisely measure the degree of anomaly based on the conformity between the model and test data
  - MDC models transform the data into probabilistic estimations
  - Proper estimation of anomaly score on these probabilities will let us correctly identify the anomalous clinical actions
- Caveat: Blindly picking the minimum probability will not satisfy our needs; E.g., prescriptions with alternative drugs
- Solutions

	Quantities Involved in Scoring	Scoring Scheme
Univariate Approach	$P(\mathbf{y}\mid\mathbf{x})$	· The complementary probability $Score_1 = 1 - P(\mathbf{y}\mid\mathbf{x})$ · Rank percentile of the probability $Score_2 = Rank[P(\mathbf{y}\mid\mathbf{x})] / N_{test}$
Multivariate Approach	$P(y_i\mid\mathbf{x})$ ** We denote $\boldsymbol{\phi}$ $= \{P(y_i\mid\mathbf{x}) : i{=}1,\ldots,d\}$	· Robust Mahalanobis Distance [Rousseeuw and Zomeren '90] $Score_3 = rd(P(y_i\mid\mathbf{x}) : i = 1,\ldots,d)$ $= (\boldsymbol{\phi} - \boldsymbol{\mu})' M^{-1} (\boldsymbol{\phi} - \boldsymbol{\mu})$ · M: minimum covariance determinant (MCD) $\boldsymbol{\mu}$: mean of $\boldsymbol{\phi} = \{P(y_i\mid\mathbf{x}) : i{=}1,\ldots,d\}$ over test data · L_x norms (L_1, L_2, L_{max}) $Score_4 = \lVert 1 - \boldsymbol{\phi} \rVert_1$ $Score_5 = \lVert 1 - \boldsymbol{\phi} \rVert_2$ $Score_6 = \lVert 1 - \boldsymbol{\phi} \rVert_{max}$
Multivariate Conditional Approach	$P(y_i\mid\mathbf{x})$, $\mathbf{x}$	· One-class SVM [Schölkopf et al. '99] · Support Vector Data Description [Tax and Duin '04]

- Using these schemes as basic building blocks, we are working on new anomaly scoring techniques

Experimental results

- **Data: Progress notes obtained from Cincinnati Children's Hospital Medical Center** [Pestian et al. '07]
  - 978 Instances (patients)
  - X: 1,449 features; Freehand notes in the bag-of-words representation
  - Y: 45 binary classes; Indicating the diseases diagnosed

- **Compared methods**:
  1) Modified Classifier Chain + Robust Mahalanobis (CC.mod+RDist)
  2) Conditional Tree BN + Robust Mahalanobis (CTBN+RDist)
  3) Binary relevance + complementary probability (BR+comP)
- 10-fold cross validation; On each round, 15% of randomly selected test data are perturbed (anomalies) by flipping 1-5 class labels
  - Anomalies represent mistaken diagnoses
- **Metric**: Area under an ROC curve (AUC)