**Introduction**

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

**Objective**

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

**Phase 1: Multi-dimensional Data Modeling of Clinical Records**

- **Graphical Representation** (e.g., \( d = 4 \))
- **Mathematical Representation**
  \[ P(Y|X) = \prod_{i=1}^{d} P(y_i|x_i, \pi(x_i)) \]
  \[ Y = \{1, \ldots, d\}; \pi(x_i) = \text{all preceding labels (tree)} \]
- **Graphical Structure**
  - *Structure learning is not required (fast)*
  - *Optimal tree structures are learned efficiently*
  - *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 limited to a tree structure*
  - *Learning the optimal structure is NP-hard*
  - *A greedy approx. is used*

**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 estimate 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**
  - *The complementary probability*
    \[ Score_{2} = 1 - P(y|x) \]
  - *Rank percentile of the probability*
    \[ Score_{2} = \text{Rank}(P(y|x)) / N_{test} \]
- **Multivariate Approach**
  - *Robust Mahalanobis Distance* (Rousseeuw and Durenre, 90)
    \[ Score_{C} = \sqrt{[\text{MCD}(x)]} \]
  - *M. minimum covariance determinant (MCD) mean of \( \phi \) (\( \phi \) = \{\phi_1, \ldots, \phi_d\}) over test data*
    \[ Score_{C} = \text{Rank}(\phi) \] \[ Score_{C} = \text{Rank}(\phi) \]
  - *Core Vector (Duin et al. 97)*
    \[ Score_{C} = \text{Rank}(\phi) \]
- **Multivariate Conditional Approach**
  - *One-class SVM* (Ras and Dua 04)
  - *Support Vector Data Description* (Tax and Duin 04)
  - Using these schemes as basic building blocks, we are working on new anomaly scoring techniques

**Quantities Involved in Scoring**

**Scoring Scheme**

**Univariate Approach**

- The complementary probability
  \[ Score_{1} = 1 - P(y|x) \]
- Rank percentile of the probability
  \[ Score_{2} = \text{Rank}(P(y|x)) / N_{test} \]

**Multivariate Approach**

- *Robust Mahalanobis Distance* (Rousseeuw and Durenre, 90)
  \[ Score_{C} = \sqrt{[\text{MCD}(x)]} \]
  - *M. minimum covariance determinant (MCD) mean of \( \phi \) (\( \phi \) = \{\phi_1, \ldots, \phi_d\}) over test data*
  \[ Score_{C} = \text{Rank}(\phi) \]
- *Core Vector (Duin et al. 97)*
  \[ Score_{C} = \text{Rank}(\phi) \]
- *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 (Poislpan 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)