Multivariate Conditional Anomaly Detection and Its Clinical Application

Charmgil Hong **Milos Hauskrecht**

{charmgil, milos}@cs.pitt.edu

Department of Computer Science
University of Pittsburgh

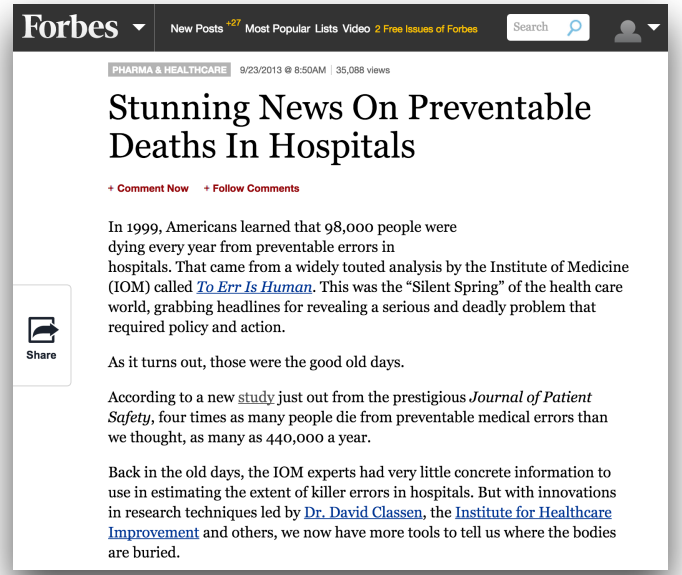
Prepared for the Twentieth AAAI/SIGAI Doctoral Consortium

Agenda

- Motivation
- Our Approach
  - Phase 1: Multi-dimensional Data Modeling
  - Phase 2: Model-based Anomaly Detection
- Conclusion

Motivation

- Reports from medical/clinical surveys
  - The occurrence of medical errors remains a persistent and critical problem
  - *Medical errors that correspond to preventable adverse events are estimated to be up to 440k patients each year* [James 2013]
    - This is *the third leading cause of death in America*



Forbes | New Posts +27 | Most Popular | Lists | Video | 2 Free Issues of Forbes | Search

PHARMA & HEALTHCARE 9/23/2013 @ 8:50AM | 35,088 views

Stunning News On Preventable Deaths In Hospitals

+ Comment Now + Follow Comments

In 1999, Americans learned that 98,000 people were dying every year from preventable errors in hospitals. That came from a widely touted analysis by the Institute of Medicine (IOM) called *To Err Is Human*. This was the "Silent Spring" of the health care world, grabbing headlines for revealing a serious and deadly problem that required policy and action.

Share

As it turns out, those were the good old days.

According to a new study just out from the prestigious *Journal of Patient Safety*, four times as many people die from preventable medical errors than we thought, as many as 440,000 a year.

Back in the old days, the IOM experts had very little concrete information to use in estimating the extent of killer errors in hospitals. But with innovations in research techniques led by Dr. David Classen, the Institute for Healthcare Improvement and others, we now have more tools to tell us where the bodies are buried.

HOSPITAL SAFETY SCORE

Home | What is Patient Safety? | Your Hospital's Safety Score | What You Can Do to Stay Safe | For Hos

How Safe is Your Hospital? Search By City/State City

I accept the Terms of Use

< Back to the newsroom

Hospital Errors are the Third Leading Cause of Death in U.S., and New Hospital Safety Scores Show Improvements Are Too Slow

Washington, D.C., October 23, 2013 – New research estimates up to 440,000 Americans are dying annually from preventable hospital errors. This puts medical errors as the third leading cause of death in the United States, underscoring the need for patients to protect themselves and their families from harm, and for hospitals to make patient safety a priority.

Captured from: http://www.forbes.com/sites/leahbinder/2013/09/23/stunning-news-on-preventable-deaths-in-hospitals/ (left) and http://www.hospitalsafetyscore.org/newsroom/display/hospitalerrors-thirdleading-causeofdeathinus-improvementstooslow (right)

Motivation

- Computer-based approaches to support clinical decisions

(1) Knowledge-driven approach

- Based on the rules or decision structures that are manually designed by human experts
  - E.g., Liver disorder diagnosis network [Onisko et al. 1999]
- Expensive to build and maintain
- Coverages are often incomplete

Motivation

- Computer-based approaches to support clinical decisions

(2) Data-driven approach

- An application of data mining and statistical machine learning techniques
- Based on the rules or decision structures that are automatically built by algorithms
- More affordable to build and maintain
- Coverages can be continuously improved along with the availability of data and techniques

Our Goal

- We aim at developing a clinical decision support system that can automatically detect erroneous clinical actions
  - Cases requiring clinical attention for reconsideration could be identified by detecting statistical anomalies in patient care patterns [Hauskrecht et al. 2007, 2013]
  - We want to identify clinical decisions that do not conform with past records
  - Virtually every hospital runs its own electronic medical record (EMR) system, to which our system can be applied

Our Approach

- A 2-phase approach
  - Phase 1: Multi-dimensional data modeling
    - We model the clinical data stored in electronic medical record (EMR) systems
  - Phase 2: Model-based anomaly detection
    - Using the model obtained in phase 1, we identify possibly erroneous clinical decisions and actions

Phase 1: Multi-dimensional data modeling

- Setting: We are given a collection of EMRs $D = \{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}\}_{n=1}^{N}$
  - A feature vector $\mathbf{x}^{(n)} = (x_1^{(n)}, \ldots, x_m^{(n)})$ of m *continuous* values that represents an observation (patient condition)
  - A decision vector $\mathbf{y}^{(n)} = (y_1^{(n)}, \ldots, y_d^{(n)})$ of d *discrete* values that represents the clinical decisions made on $\mathbf{x}^{(n)}$
    - For simplicity, this presentation will focus only on the *binary* decision cases
- Objective: We want to accurately and efficiently learn a compact model of complex clinical data
- Challenge: both $\mathbf{x}$ and $\mathbf{y}$ are high-dimensional

Phase 1: Multi-dimensional data modeling

- The *multi-dimensional classification (MDC)* problem formulates this kind of modeling situations [Zhang and Zhou 2013]
  - In MDC, we want to learn a function that assigns to each observation (patient), represented by its feature vector $\mathbf{x}$, the most probable assignment of the decisions (clinical actions) $\mathbf{y}$
  - Assuming the 0-1 loss function, the optimal function h^* maps an observation to the maximum a posterior (MAP) assignment of the decisions

$$
\begin{aligned}
h^*(\mathbf{x}) &= \arg\max_{\mathbf{y}} P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) \\
&= \arg\max_{y_1, \ldots, y_d} P(Y_1 = y_1, \ldots, Y_d = y_d | \mathbf{X} = \mathbf{x})
\end{aligned}
$$

A Simple MDC Solution: d Independent Models

- Idea [Clare and King 2001; Boutell et al. 2004]
  - Transform an MDC problem to multiple single-label classification problems
  - Learn d independent classifiers for d decision variables
- Illustration

D_{train}	X_1	X_2	Y_1	Y_2	Y_3
$n=1$	0.7	0.4	1	1	0
$n=2$	0.6	0.2	1	1	0
$n=3$	0.1	0.9	0	0	1
$n=4$	0.3	0.1	0	0	0
$n=5$	0.8	0.9	1	0	1

$h_1 : \boldsymbol{X} \rightarrow Y_1$

$h_2 : \boldsymbol{X} \rightarrow Y_2$

$h_3 : \boldsymbol{X} \rightarrow Y_3$

A Simple MDC Solution: d Independent Models

- Advantage
  - Computationally very efficient
- Disadvantage
  - Not suitable for our objective
    - Does not find the most probable assignment
      - Instead, it maximizes the marginal distribution of each decision variable
    - Does not capture the correlations among the decision variables
      - Clinical decisions often show correlations
        - E.g., a set of medications in relations

Examples: Correlations in Clinical Decisions

- A set of medications in relations
  - Medications that are usually prescribed together
  - Alternative medications that only one of them is prescribed
  - Adverse medications that should not be prescribed together

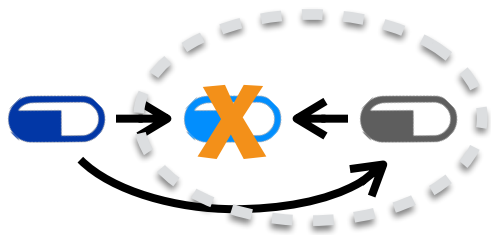
Examples: Correlations in Clinical Decisions

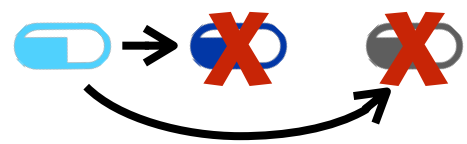
- Correlations among medications

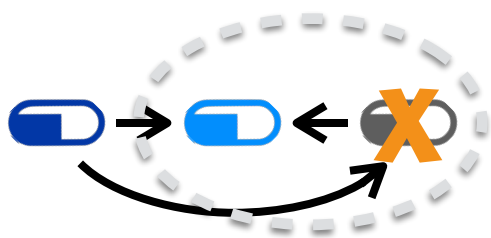
Medications usually given together

Alternative medications among which only one is given

Adverse medications should not be given together



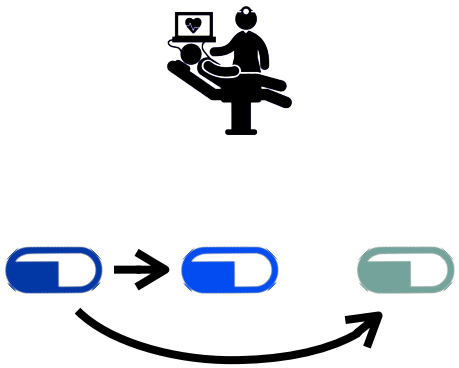




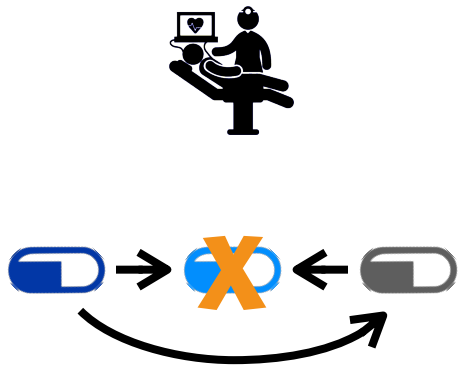
Examples: Correlations in Clinical Decisions

- Correlations among medications

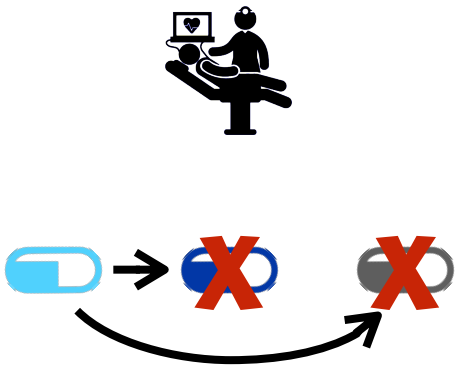
Medications usually given together

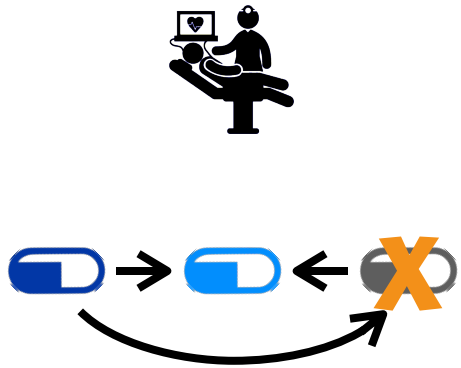


Alternative medications among which only one is given



Adverse medications should not be given together

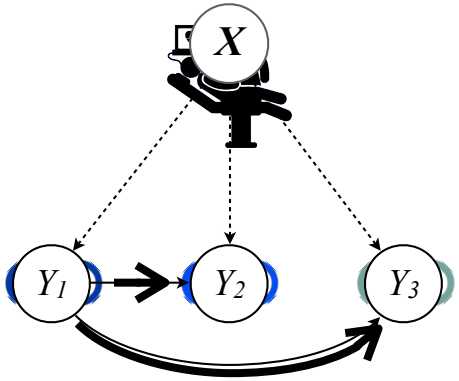




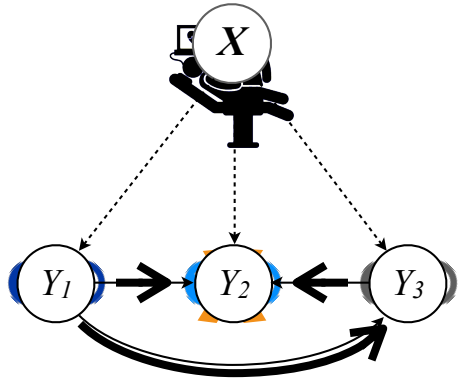
Examples: Correlations in Clinical Decisions

- Correlations among medications

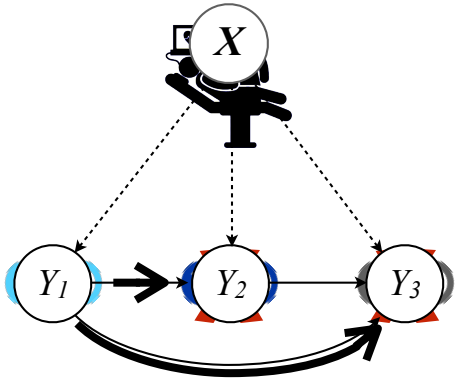
Medications usually given together

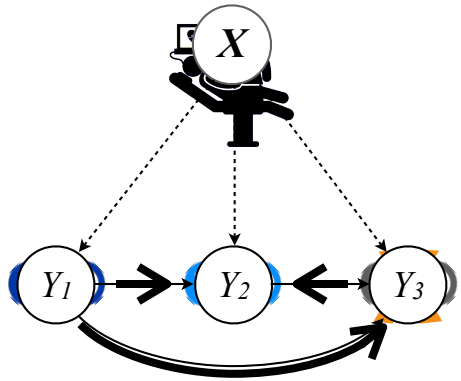


Alternative medications among which only one is given



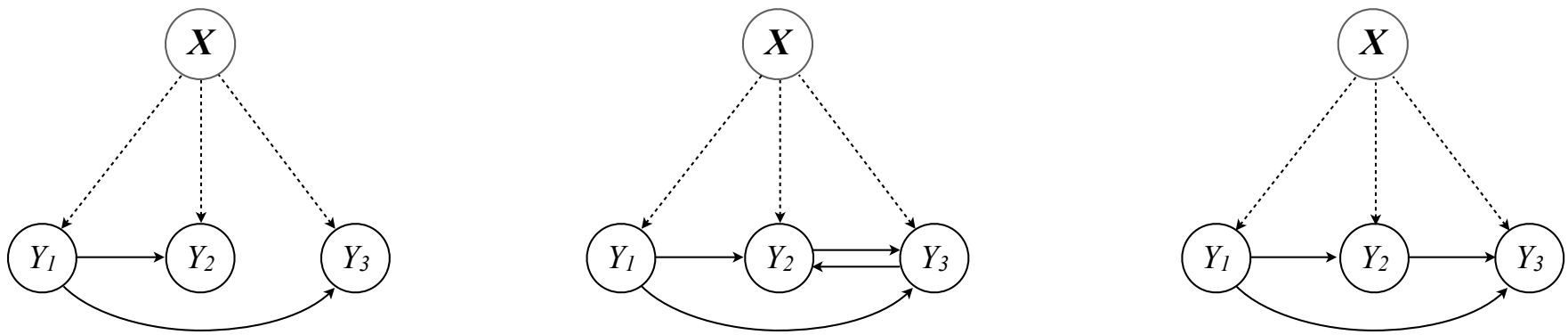
Adverse medications should not be given together





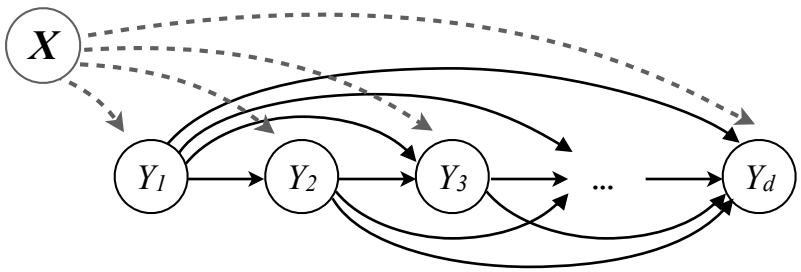
Examples: Correlations in Clinical Decisions

- Learning the correlation structure in clinical decisions is the key to facilitate the clinical data modeling!



Learning Correlations in Multiple Decisions with CC

- *Classifier Chains* (CC) [Read et al. 2009]
  - Represents the chain rule of the probability, conditioned on observations
  - On m variables of patient condition and d decision variables, CC defines the joint probability $P(Y_1, \ldots, Y_d|\mathbf{X})$ as:

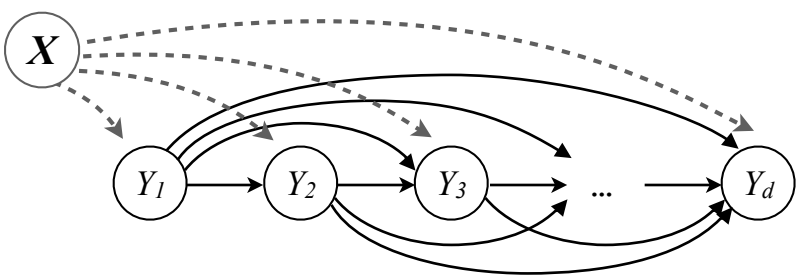


$$P(Y_1, ..., Y_d|\mathbf{X}) = \prod_{i=1}^{d} P(Y_i|\mathbf{X}, Y_1, ..., Y_{i-1})$$

$$= P(Y_1|\mathbf{X}) \cdot P(Y_2|\mathbf{X}, Y_1) \cdot ... \cdot P(Y_d|\mathbf{X}, Y_1, ..., Y_{d-1})$$

Learning of Multiple Decisions with CC

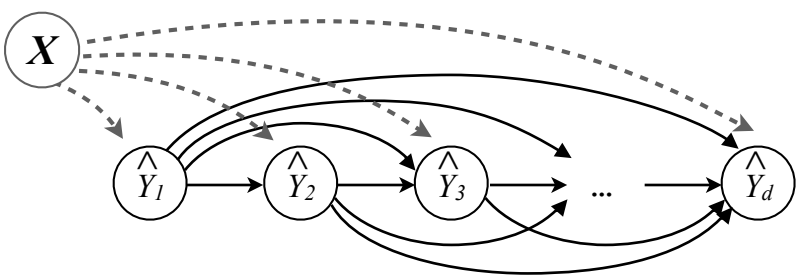
- Learning of CC
  - Using the decomposition along the "chain," the distribution of each decision Y_i is modeled using a probabilistic function (e.g., logistic regression)



$$P(\mathbf{Y}|\mathbf{X}) =$$

Prediction of Multiple Decisions with CC

- Prediction with CC
  - Make a prediction on each decision variable Y_i along the chain order; use the predictions of the preceding decisions as observations (in addition to $\mathbf{x}$) for the following chains



$$P(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{d} P(Y_i|\mathbf{X}, Y_1, ..., Y_{i-1})$$
$$= P(Y_1|\mathbf{X}) \cdot P(Y_2|\mathbf{X}, Y_1) \cdot ... \cdot P(Y_d|\mathbf{X}, Y_1, ..., Y_{d-1})$$

Q: What if a prediction is wrong? ***Error propagates***

Q: Does $\mathbf{X}$ have the same predictability towards $Y_1, \ldots Y_d$?

Chain order matters

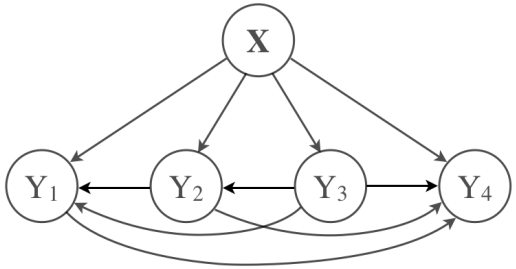
Contribution 1: Algorithmic enhancement [Hong et al. 2015]

- An issue with CC
  - The order in $\{Y_1, \ldots, Y_d\}$ actually affects the model and prediction accuracy
  - Knowing a proper ordering of chain is desired
  - However, the size of structure space is extremely large ($d!$)
- Solution: ***CC.algo***
  - A greedy structure learning algorithm that picks the chain order
  - Performs very well in practice

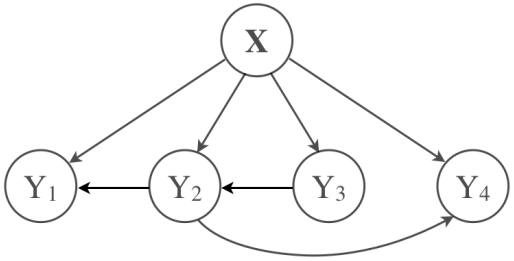
Contribution 2: Structural modification [Batal et al. 2013]

- An issue with CC
  - CC does not provide "optimal structure" learning
  - Greedy prediction algorithm does not produce the exact MAP assignment
    - The exact MAP assignment on CC takes exponential in d time [Dembczynski et al. 2010]
- Solution: ***CC.tree***
  - Restrict the correlation structure to be a *tree*

*An example **CC** (d=4)*



*An example **CC.tree** (d=4)*



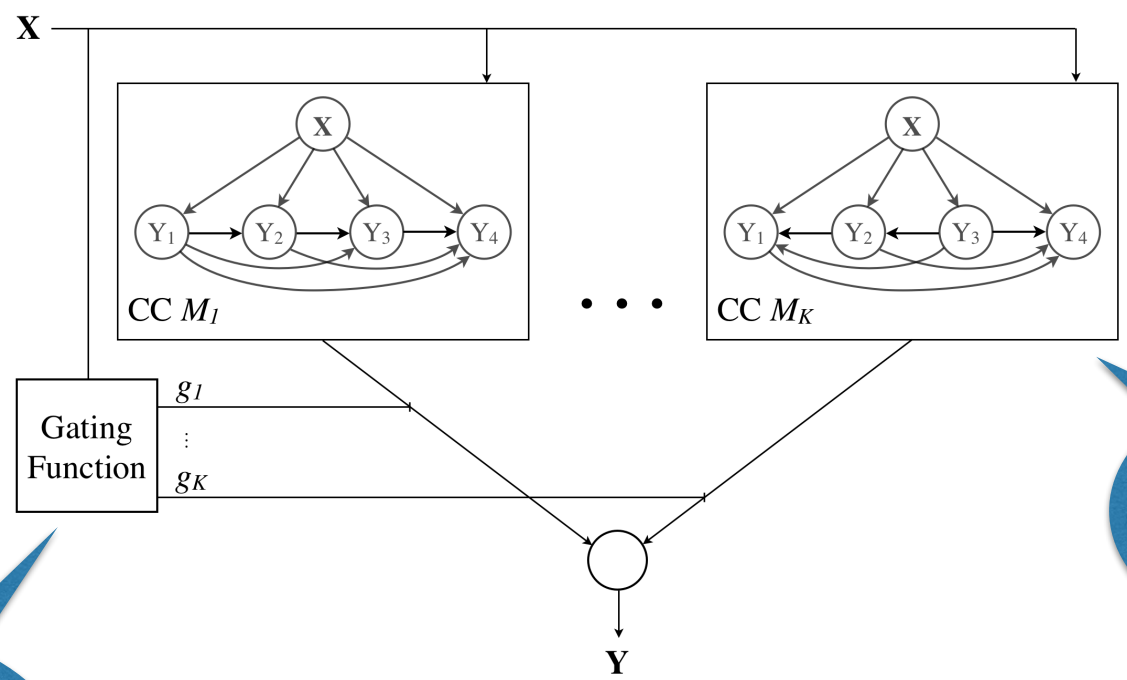
Contribution 3: Mixture extensions [Hong et al. 2014, 2015]

- An issue with CC
  - CC cannot fully recover the joint distribution $P(Y_1,\ldots,Y_d|\mathbf{X})$ in practice
  - The mixture approaches let us learn multiple CCs and combine them to produce more accurate outputs
- Solution: ***CC.me***
  - We extended the *mixtures-of-experts* [Jacobs et al. 1991] framework to solve the MDC problem
  - Our extension manages multiple correlation structures and produces more accurate data models

Contribution 3: Mixture extensions [Hong et al. 2014, 2015]

- Solution: ***CC.me***

An example CC.me (d=4)



Multiple CC models

Input (X) dependent weighting

Phase 1: Experimental Results

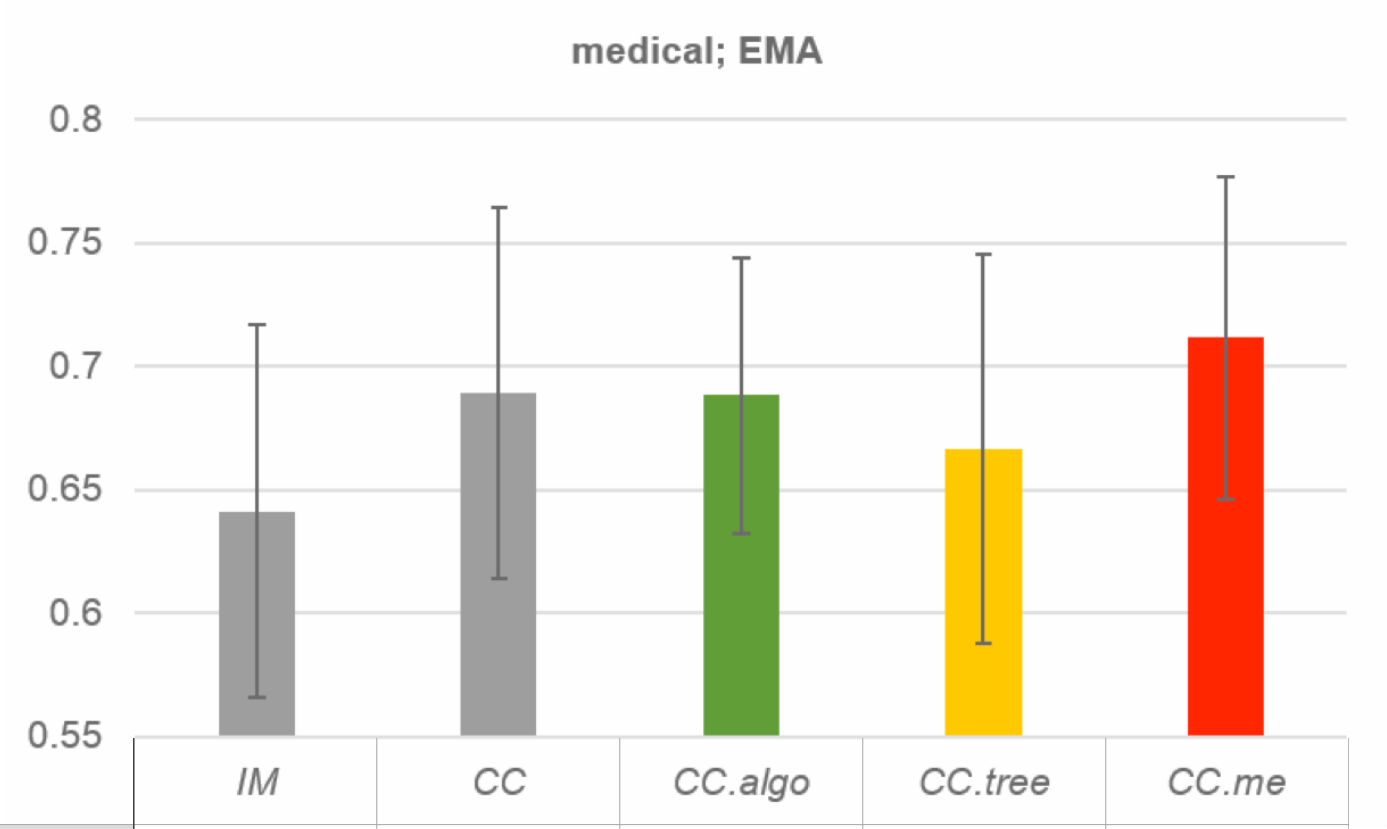
- Compared methods
  - Independent Models (*IM*) — *baseline*
  - Classifier Chains (*CC*) — *baseline*
  - Algorithmic extension (*CC.algo*)
  - Structural extension (*CC.tree*)
  - Mixtures-of-Experts extension (*CC.me*)

Phase 1: Experimental Results

- Data: Progress notes obtained from Cincinnati Children's Hospital Medical Center [Pestian et al. 2007]
  - 978 patient records
  - $\mathbf{X}$: 1,449 features; Freehand notes in the bag-of-words representation
  - $\mathbf{Y}$: 45 binary classes; Indicating the diseases diagnosed
- Metrics
  - Exact match accuracy (EMA): the probability of all decisions are predicted correctly
  - Conditional log-likelihood loss (CLL-loss): shows the model fitness to the test data
    - the sum of negative log-probability on test data given a trained model

Phase 1: Experimental Results

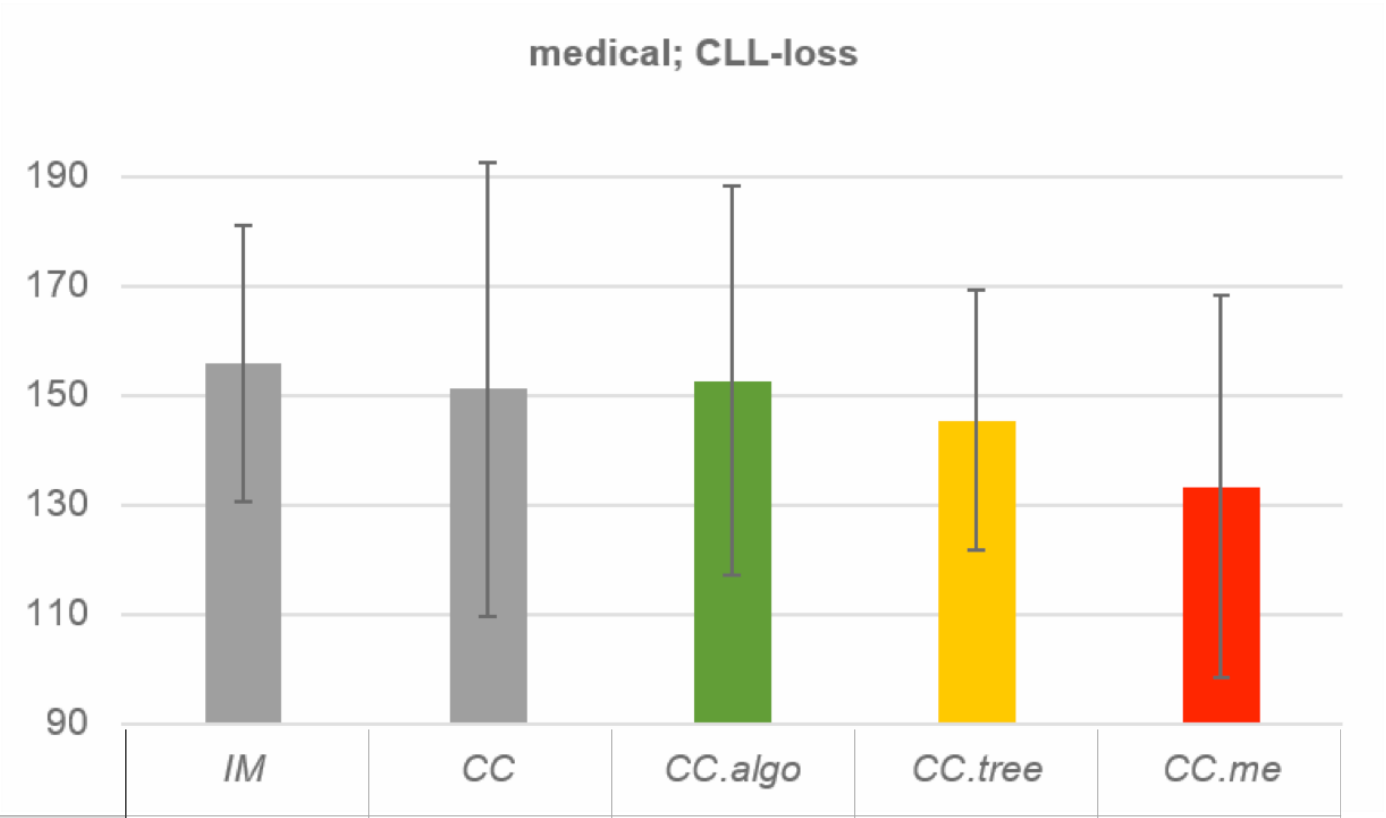
- Exact match accuracy (EMA; *higher is better*)



	IM	CC	CC.algo	CC.tree	CC.me
EMA	0.64 ± 0.08	0.69 ± 0.08	0.69 ± 0.06	0.67 ± 0.08	0.71 ± 0.07
Rank (paired t-test α = 0.05)	5	2	2	4	1

Phase 1: Experimental Results

- Conditional log-likelihood loss (CLL-loss; *smaller is better*)



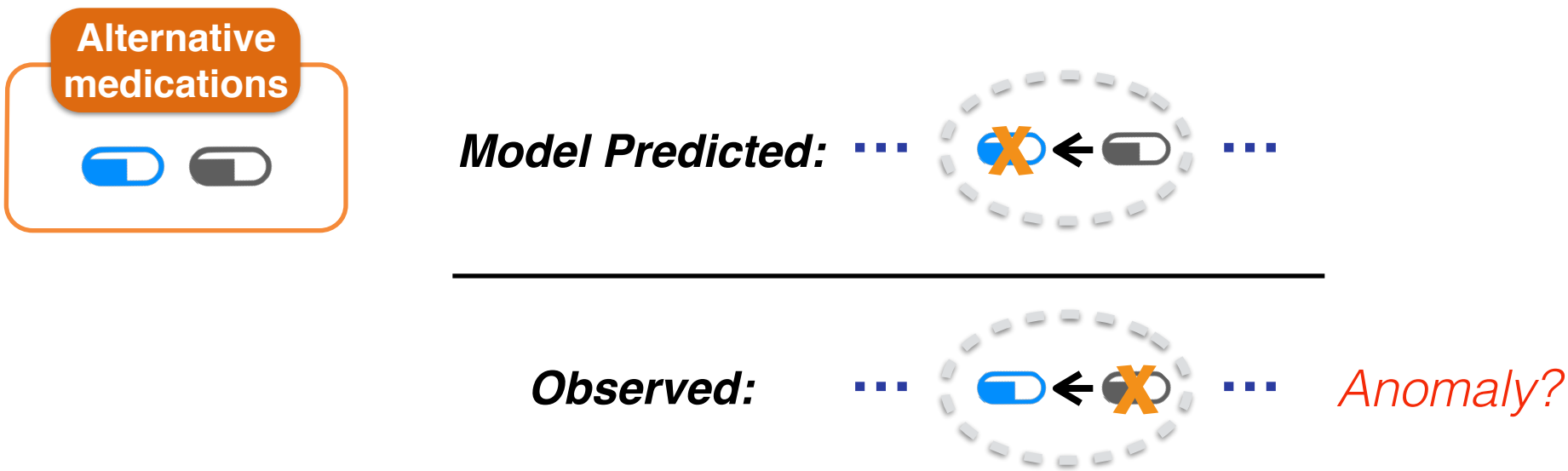
	IM	CC	CC.algo	CC.tree	CC.me
CLL-loss	155.9 ± 25.2	151.1 ± 41.4	152.7 ± 35.6	145.4 ± 23.8	133.3 ± 34.8
Rank (paired t-test α = 0.05)	3	3	3	2	1

Phase 2: Model-based anomaly detection

- Setting
  - We are given a trained model M (using any of models from phase 1) and a set of unseen test data $D_{test} = \{\mathbf{x}^{(l)}, \mathbf{y}^{(l)}\}_{l=1}^{L}$ which may include anomalous clinical decisions
- Objective
  - We want to identify anomalous observations-decisions pairs in D_{test} using M

How to properly measure the anomalousness?

- Conventional model-based approach: *univariate anomaly scoring scheme* [Filzmoser et al. 2006]
  - Simply consider the joint likelihood $P(\mathbf{y}|\mathbf{x}; M)$
  - The complementary probability $1 - P(\mathbf{y}|\mathbf{x}; M)$ indicates the degree of anomalousness of decisions $\mathbf{y}$ on patient $\mathbf{x}$



Our Approach to Score Anomaly

- Our approach: *multivariate anomaly scoring scheme*
  - Given a trained model M and test data $D_{test} = \{\mathbf{x}^{(l)}, \mathbf{y}^{(l)}\}_{l=1}^{L}$

(1) Transform the observations-decisions pairs into a vector of probabilistic estimation $\boldsymbol{\phi}^{(l)} = (P(y_1^{(l)}|\mathbf{x}^{(l)}; M), \ldots, P(y_d^{(l)}|\mathbf{x}^{(l)}; M))$

(2) Properly measure the *anomaly score* using $\boldsymbol{\phi}^{(l)}$

Multivariate Anomaly Scoring

- Consider the likelihood $\boldsymbol{\phi}^{(l)} = (P(y_1{}^{(l)}|\mathbf{x}^{(l)}; M), \ldots, P(y_d{}^{(l)}|\mathbf{x}^{(l)}; M))_{l=1}^{L}$ on every decision dimension to score anomaly
  - *Scoring example*: Using the robust distance [Rousseeuw and Zomeren '90]
    - $Score_{rd}(\boldsymbol{\phi}^{(l)}) = (\boldsymbol{\phi}^{(l)} - \boldsymbol{\mu})' M^{-1} (\boldsymbol{\phi}^{(l)} - \boldsymbol{\mu})$

      where M: minimum covariance determinant (MCD)

      $\boldsymbol{\mu}$: mean of $\boldsymbol{\phi} = (P(y_i|\mathbf{x}) : i = 1, \ldots, d)$ over test data
    - A variant of the Mahalanobis distance

Preliminary Results

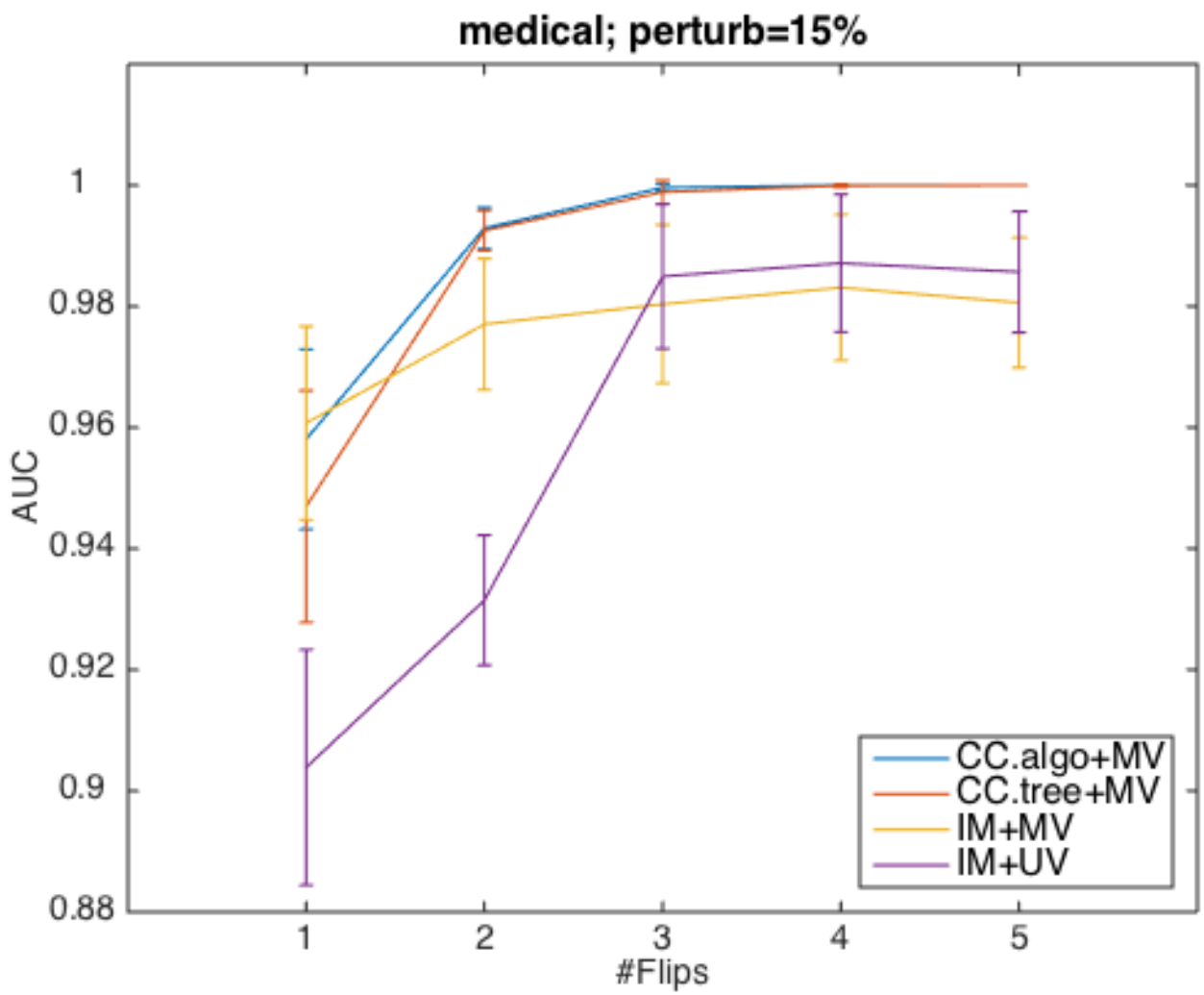
- Task: To identify incorrect disease diagnoses
- Data: Progress notes obtained from Cincinnati Children's Hospital Medical Center [Pestian et al. 2007]
  - 978 patient records
  - $\mathbf{X}$: 1,449 features; Freehand notes in the bag-of-words representation
  - $\mathbf{Y}$: 45 binary classes; Indicating the diseases diagnosed

Preliminary Results

- Experiment
  - Compared methods
    - *CC.algo* [Hong et al. 2015] + *Robust Distance (CC.algo+MV)*
    - *CC.tree* [Batal et al. 2013] + *Robust Distance (CC.tree+MV)*
    - *Independent model* [Clare and King 2001; Boutell et al. 2004] + *Robust Distance (IM +MV)*
    - *Independent model* [Clare and King 2001; Boutell et al. 2004] + Complementary Probability *(IM+UV)*
  - 10-fold cross validation; on each round, 15% of anomalies are injected to the test set by flipping 1-5 decisions
- Metric: Area under receiver operating characteristic (AUC)

Preliminary Results

- Area under receiver operating characteristic (AUC; *higher is better*)



Multivariate Anomaly Scoring

- This part is in progress
- We are trying to better understand about the space of the conditional likelihood estimate $\boldsymbol{\phi} = (P(y_1|\mathbf{x}; M), \ldots, P(y_d|\mathbf{x}; M))$
- Future work
  - Developing robust anomaly scoring schemes that have reasonable semantics
  - Identifying the root causes of anomalies
  - Unifying the phase 1 and 2 into a single optimization formulation

Conclusion

- We are aiming at building clinical decision support systems by detecting anomalies in clinical records
  - We first model the past clinical data stored in EMRs
  - We then use the model to identify anomalies that contains the clinical decisions that do not conform with past records
- Clinical data modeling:
  - We developed and improved multi-dimensional data models and methods
- Anomaly detection:
  - We proposed a new approach to multivariate anomaly detection that estimates the anomalousness of observations-decisions pairs, using the conditional likelihood under a trained model

Acknowledgement

- This work was supported by grants R01LM010019 and R01GM088224 from the NIH. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

- Charmgil thanks to all the colleagues are/were in 210 S Bouquet St. #5406, Pittsburgh, PA 15213:

  - Dr. Milos Hauskrecht, Dr. Iyad Batal, Dr. Saeed Amizadeh, Dr. Quang Nguyen, Zitao Liu, Eric Heim, Mahdi Pakdaman, Salim Malakouti, Jeongmin Lee, Zhipeng Luo

References

- [James 2013] J. T. James. A new, evidence-based estimate of patient harms associated with hospital care. Journal of patient safety, 9(3):122–128, Sept. 2013.
- [Onisko et al. 1999] A. Oniśko, M.J. Druzdzel, H. Wasyluk, A Bayesian network model for diagnosis of liver disorders, in: Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering, vol. 2, Warsaw, Poland, December 2–4, 1999, pp. 842–846
- [Hauskrecht et al. 2007] M. Hauskrecht, M. Valko, B. Kveton, S. Visweswaram, and G. Cooper. 2007. Evidence-based anomaly detection. In Annual American Medical Informatics Association Symposium,319–324.
- [Hauskrecht et al. 2013] M. Hauskrecht, I. Batal, M. Valko, S. Visweswaran, G. F. Cooper, and G. Clermont. Outlier detection for patient monitoring and alerting. Journal of Biomedical Informatics, 46(1):47–55, Feb. 2013.
- [Zhang and Zhou] M. Zhang and Z. Zhou. A review on multi-label learning algorithms. Knowledge and Data Engineering, IEEE Transactions on, PP(99):1, 2013.
- [Clare and King 2001] A. Clare and R. D. King. Knowledge discovery in multi-label phenotype data. In In: Lecture Notes in Computer Science, pages 42–53. Springer, 2001.
- [Boutell et al. 2004] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. Pattern Recognition, 37(9):1757 – 1771, 2004.
- [Read el al. 2009] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Clas- sifier chains for multi-label classification. In Proceed- ings of the European Conference on Machine Learn- ing and Knowledge Discovery in Databases. Springer- Verlag, 2009.
- [Dembczynski et al. 2010] K. Dembczynski, W. Cheng, and E. Hu¨llermeier. Bayes optimal multilabel classification via probabilistic clas- sifier chains. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 279–286. Omnipress, 2010.

References

- [Hong et al. 2015] C. Hong, I. Batal, and M. Hauskrecht. A general- ized mixture framework for multi-label classification. In Proceedings of the 2015 SIAM International Con- ference on Data Mining. SIAM, 2015.
- [Batal et al. 2013] I. Batal, C. Hong, and M. Hauskrecht. An efficient probabilistic framework for multi-dimensional classifi- cation. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Manage- ment, CIKM '13, pages 2417–2422. ACM, 2013.
- [Hong et al. 2014] C. Hong, I. Batal, and M. Hauskrecht. A mixtures- of-trees framework for multi-label classification. In Proceedings of the 23nd ACM International Conference on Information and Knowledge Management, CIKM '14. ACM, 2014.
- [Kumar et al. 2012] A. Kumar, S. Vembu, A. K. Menon, and C. Elkan. Learning and inference in probabilistic classifier chains with beam search. In Proceedings of the 2012 Euro- pean Conference on Machine Learning and Knowledge Discovery in Databases. Springer-Verlag, 2012.
- [Meila and Jordan 2001] M. Meila, M. Jordan. Learning with Mixtures of Trees. Journal of Machine Learning Research, 1(Oct):1-48, 2000.
- [Jacobs et al. 1991] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. Neural Comput., 3(1):79–87, Mar. 1991.
- [Pestian et al. 2007] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hov- ermale, N. Johnson, K. B. Cohen, and W. Duch. A shared task involving multi-label classification of clin- ical free text. In Proceedings of the Workshop on BioNLP 2007, pages 97–104, 2007.
- [Rousseeuw and Zomeren '90] P. J. Rousseeuw and B. C. v. Zomeren. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85(411):pp. 633–639, 1990.

Thanks!

Multivariate Conditional Anomaly Detection and Its Clinical Application

Point of Contact: Charmgil Hong
www.cs.pitt.edu/~charmgil
charmgil@cs.pitt.edu