Assisting Clinical Decision Making: Perioperative Risk Prediction

Hanyang Liu
CSE 531A Lecture

Slides: Bing Xue, PhD
Perioperative Handoff

- **OR-ICU handoff**: Transfer patients from operation room to ICU room
- **Intraoperative handoff**: Change of shifts during surgery
- **Handoff lasts only for a very short time**
  - must be very informative
  - concentrate on key information relevant to risk mitigation

Early risk identification  Modifiable risks
Why is Early Prediction Important?

- Surgeries are highly risky:
  - $\geq 10\%$ of surgical patients experience major postoperative (postop) complications

- Early prediction is important
  - Patients: early intervention and improved outcomes
  - Hospitals: better resource allocation, reduced cost
  - Clinicians: better quality of care
Modifiable risks

Modifiable risks

Potentially preventable/improvable through early detection and intervention

Modifiable postop complications – our targets to predict

- acute kidney injury (AKI)
- delirium
- deep vein thrombosis (DVT)
- pulmonary embolism (PE)
- pneumonia.
Risk prediction with ML

- **Inputs** - Electronic Health Records (EHR):
  - Collected different times: *preoperative* (preop) & *intraoperative* (intraop)
  - Multi-modal input:
    - **Static Features**: demographics, comorbidities, lab tests, treatment/interventions, drugs, etc.
    - **Time Series**: vital signs, events, measurements, etc.
    - **Texts**: Clinical notes.

- **Outputs**: probability of post op (future) complications
Overall Framework

- NLP Method
- Time Series Method
- Feature Fusion
- ML Predictor
**What are clinical texts?**

- Clinical texts have complex abbreviations or medical terminologies
- Clinical texts can be structured or unstructured

<table>
<thead>
<tr>
<th>AN_PROC_NAME</th>
<th>AN_DATE</th>
<th>PROCED_ICD</th>
<th>PROCED_ICD_SHORT_TITLE</th>
<th>PROCED_ICD_LONG_TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOPHAGEOGASTRODUODENOSCOPY wu prep (N/A); CO...</td>
<td>2019-03-05 00:00:00</td>
<td>[9604, 9671, 9634]</td>
<td>Insert endotracheal tube, 'Gastric tube irri...</td>
<td>Insertion of endotracheal tube Other irrigatio...</td>
</tr>
<tr>
<td>BILATERAL MYRINGOTOMY TUBE INSERTION (Bilatera...</td>
<td>2019-03-14 00:00:00</td>
<td>[3613, 3615, 3961, 8872, 9904, 9905, 9907]</td>
<td>Aortocoronary bypass-3 cor art, '1 int mam-cor ar...</td>
<td>Aortocoronary bypass of three coronary artery...</td>
</tr>
<tr>
<td>FUSION SPINAL ANTERIOR LUMBAR/THORACIC WITH IN...</td>
<td>2019-03-28 00:00:00</td>
<td>[151]</td>
<td>[Ex cereb meningial les]</td>
<td>Excision of lesion or tissue of cerebral meninges</td>
</tr>
<tr>
<td>EXCHANGE STENT - URERAL (Right Ureter)</td>
<td>2019-03-05 00:00:00</td>
<td>[3601, 3607, 3723, 8852, 8856, 9920]</td>
<td>[Ins drug-elut corony st, Rt/left heart ca...</td>
<td>Insertion of drugeluting coronary artery stent...</td>
</tr>
<tr>
<td>Robotic Assisted Laparoscopic Partial Fundopi...</td>
<td>2019-03-25 00:00:00</td>
<td>[3601, 3607, 3722, 8856, 9920, 3722, 8856, 885...</td>
<td>[Ins drug-elut corony st, Left heart cardi...</td>
<td>Insertion of drugeluting coronary artery stent...</td>
</tr>
</tbody>
</table>
How to read the texts?

- **Word Embedding**: Convert words to embeddings
How to read the texts?

- **Continuous Bag-of-Words (CBOW)**
  - predicts which is the most likely word in the given context

- **GloVe**
  - trained on aggregated global word to word co-occurrence matrix from a given text collection of text documents.

- **FastText**
  - It is an extension to CBOW model
  - but unlike Word2Vec which feeds whole words into the neural network, FastText first breaks the words into several sub-words (or n-grams) and then feed them into the neural network.
Limitations of Word Embedding

- Not take consider order of words in which they appear
  - leads to loss of **syntactic** and **semantic** understanding of the sentence.
  - E.g., “You are going there to teach not play” vs. “You are going there to play not teach.”

- Same word may differ in meaning in context
  - “I have scuba diving in my **bucket** list.” And “There is a **bucket** filled with drinking water.”
  - In both the sentences, the word “bucket” has different meanings.
Sentence Embedding

- A simple way of obtaining sentence embedding is by averaging the word embeddings of all the words present in the sentence. But they are not accurate enough.

- **ELMO** (Embedding from Language Model)
  - bi-directional deep LSTM network for producing vector representation. ELMo considers words within which context they have been used rather than creating dictionary of words with its vector form.

- **Doc2Vec**
  - Aka Paragraph Vector, is an extension of Word2Vec that learns vector representations of documents rather than words.
  - learns vector representations of documents by combining the word vectors with a document-level vector.
We compared the performance between various word embedding & sentence embedding methods, and found out that GloVe performs the best.

- On both BJC dataset and MIMIC-III dataset
- On various tasks: postoperative complications, mortality, LoS, discharge, etc.
Overall Framework

- **Glove**
- **Time Series Method**
- **Concatenated Features**
- **ML Predictor**
Modeling Time Series

- Recurrent Neural Networks (RNN): LSTM, GRU
  - Use RNN to learn the final hidden states from the multivariate time series

The architectures for time-series forecasting model: (a) Recurrent Neural Network (RNN); (b) Long-Short Term Memory (LSTM).
Challenges in time series

- **Granularity:**
  - Different vital signs are collected with different frequency (per minute, hour, day, etc).

- **Sequence lengths:**
  - Surgery may last very long (with more observations/data in the time series) or short (with few observations), depending on individual’s surgery.
Challenges in time series

- Even for the same patient, time series are irregular:
  - Each feature is collected with different irregularity
Challenges in EHR data: prevalence of missing values

- How to characterize patients with missing values:
  - The correlation between input variables is undermined by missing values;
    - Example: various missing rates of static features in different datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Delirium</th>
<th>OR</th>
<th>MIMIC-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Rate</td>
<td>52.65%</td>
<td>2.50%</td>
<td>11.79%</td>
</tr>
</tbody>
</table>
There are 2 problems to solve:

- How to deal with the *irregularity*
- How to capture both the local and global patterns

What are the candidate solutions?

- Feature engineering w/ rule-based imputation
- Learning to impute
- Learning to impute/predict jointly
Solution 1

- Feature imputation for time series
  - Zero filling
  - Mean imputation
  - Carry forward
- Summary feature extraction
  - E.g., max, std, entropy, kurtosis, skewness of a sequence
- Capturing missingness (irregularity) patterns
  - Concatenate with mask (indicating missingness)
Solution 2

- Learning to impute
  - MissForest
    - find the "best" predictions for the missing values given some set of predictors and random forest predictor
  - MICE
    - runs multiple imputation models on bootstrapped samples of observed data and randomly selects the predicted value from one of the models.
Solution 3

- Learning to impute/predict jointly
- Example A: BRITS (Bidirectional Recurrent Imputation for Time Series)
  - Use a mask vector to record the missingness
  - Use the previous hidden states of RNN to predict the next observation
  - If next observation is missing, then use the predicted value; otherwise use the actual observation

Figure 1: An example of multivariate time series with missing values. $x_1$ to $x_6$ are observed at $s_1...6 = 0, 2, 7, 9, 14, 15$ respectively. Considering the 2nd feature in $x_6$, the last observation of the 2nd feature took place at $s_2 = 2$, and we have that $\delta^6_2 = s_6 - s_2 = 13$.

Figure 2: Imputation with unidirectional dynamics.
Example B: Gated Recurrent Unit GRU-D

- Similar to BRITS in the setting
- Use GRU instead of RNN
- Instead of imputing observations at each time step, use the time interval since last observation to estimate the current hidden state

Figure 3: Graphical illustrations of the original GRU (left) and the proposed GRU-D (right) models.
Comparing imputation methods

- Indicator method
  - Use mean imputation but use an extra indicator vector to highlight the missingness
  - Achieved better performance than other imputation methods
    - Indicator reflects clinicians’ intention to skip/order certain lab tests, treatments, etc
    - Our findings were empirically and theoretically validated by Mike et. al (KDD’23)

<table>
<thead>
<tr>
<th>Imputation Method</th>
<th>AUROC</th>
<th>AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dummy Indicator</td>
<td>0.905 (0.903,0.907)</td>
<td>0.208 (0.203,0.213)</td>
</tr>
<tr>
<td>Mean</td>
<td>0.901 (0.898,0.904)</td>
<td>0.206 (0.199,0.213)</td>
</tr>
<tr>
<td>Median</td>
<td>0.901 (0.898,0.904)</td>
<td>0.204 (0.199,0.210)</td>
</tr>
<tr>
<td>Mode</td>
<td>0.902 (0.900,0.904)</td>
<td>0.203 (0.198,0.208)</td>
</tr>
<tr>
<td>kNN-3</td>
<td>0.884 (0.882,0.887)</td>
<td>0.182 (0.176,0.188)</td>
</tr>
<tr>
<td>kNN-5</td>
<td>0.884 (0.882,0.887)</td>
<td>0.182 (0.176,0.188)</td>
</tr>
<tr>
<td>MICE</td>
<td>0.893 (0.890,0.896)</td>
<td>0.193 (0.187,0.200)</td>
</tr>
<tr>
<td>MissForest</td>
<td>0.890 (0.888,0.893)</td>
<td>0.184 (0.178,0.189)</td>
</tr>
</tbody>
</table>
Overall Framework

- Glove
- Feature Engineering
- Concatenated Features
- ML Predictor
Postop Complication Prediction -- Dataset

111888 Total surgery cases recorded (June 1, 2012-August 31, 2016)

AKI  Delirium  DVT  Pneumonia  PE

711 Imputation and feature engineering

106870 AKI (positive ratio, 6.1%)
12919 Delirium (positive ratio, 52.6%)
111888 DVT (positive ratio, 1.3%)
111888 PE (positive ratio, 0.5%)
111888 Pneumonia (positive ratio, 2.1%)

5 Random shuffles

First iteration
Second iteration
Third iteration
Fourth iteration
Fifth iteration

Training folds  Testing fold
Observations

- We start with some existing machine learning models: logistic regression, random forest, gradient boosting tree, deep neural network, etc.
- Preoperative features are more informative than intraoperative time series!
Observations

- We start with some existing machine learning models: logistic regression, random forest, gradient boosting tree, deep neural network, etc.
  - Features with higher missing rate also (slightly) help with prediction
  - More features, (slightly) better performance
This brings up our next question

How to build models with high-dimensional inputs:

- Many of the input variables are correlated:
  - OR-ICU Handoff: >700 input features; National COVID Cohort Collaborative: >10k input features only in measurements;

- ML models need to handle large data
  - Curse of dimensionality:
    - Data sparsity
    - Training stage: slow and overfitting
    - Application stage: generalization of the model

- Objective: can we distill the intrinsic characteristics of patients into a lower dimensional representation?
Objectives

- A new representation / latent embedding:
  - Low-dimensional (easy for predictions);
  - No missingness;
  - Learns the implicit, nonlinear relationship between input features
Solution: generative modeling

Generative modeling:

- Encoder-Decoder Architecture:
  - Encoder transforms inputs into a latent representation
  - Decoder transforms latent representation into original/new representation

Why Variational Autoencoder (VAE):

- Self-supervised training:
  - VAE has a reconstruction term to self-supervise the latent representation
  - Nice properties in the latent space
Autoencoder (AE): Reconstruct from latent attributes

encoder

Latent attributes

decoder

- Smile: 0.99
- Skin tone: 0.85
- Gender: -0.73
- Beard: 0.85
- Glasses: 0.002
- Hair color: 0.68
VAE: Represent Latent Attribute As Probability Distribution
VAE: Reconstruct from samples in latent space

encoder

Latent distributions

Sample

Smile: 0.17
Skin tone: 0.28
Gender: -0.11
Beard: 0.66
Glasses: -0.14
Hair color: 0.26

decoder

We expect an accurate reconstruction for any sample from the latent state distributions
AE vs. VAE

**Simple Autoencoders**
- Input: $x$
- Encoding: $z = e(x)$
- Decoding: $d(z)$

**Variational Autoencoders**
- Input: $x$
- Encoding: $p(z|x)$
- Sampling: $z \sim p(z|x)$
- Decoding: $d(z)$
AE vs. VAE

Difference between a "regular" and an "irregular" latent space.
Learned Latent Space

PCA

AE (reconstruction)

Gaussian dist. (regularization)

VAE (reconstruction + regularization)
Clinical VAE (cVAE)

- Open challenges to address
  - Performance: latent encoding does not benefit downstream learning
  - Entangled latent space: difficult to interpret and trust the predictions

- Our proposed cVAE:
  - Supervised autoencoding: use training labels to guide the latent encoding
  - Disentanglement: each dimension represents a single aspect of information → more “interpretable”
## Evaluation: Predictive Performance

- Lower dimensionality
- Better downstream prediction
- Eliminates the needs for a predictor

### Table: Prediction performance for postoperative delirium (5-fold CV)

<table>
<thead>
<tr>
<th>Transformation Method \ (d=10)</th>
<th>Direct Prediction</th>
<th>LR</th>
<th>XGBoost</th>
<th>SVM</th>
<th>DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Precision</td>
<td>ROC AUC</td>
<td>Average Precision</td>
<td>ROC AUC</td>
<td>Average Precision</td>
</tr>
<tr>
<td>PCA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.717 (.009)</td>
<td>.739 (.015)</td>
</tr>
<tr>
<td>ICA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.747 (.009)</td>
<td>.769 (.012)</td>
</tr>
<tr>
<td>GMM</td>
<td>.720 (.007)</td>
<td>.732 (.010)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.643 (.007)</td>
<td>.657 (.015)</td>
</tr>
<tr>
<td>VAE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.647 (.007)</td>
<td>.667 (.009)</td>
</tr>
<tr>
<td>pi-VAE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.690 (.014)</td>
<td>.714 (.018)</td>
</tr>
<tr>
<td>cVAE-P</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.656 (.010)</td>
<td>.672 (.007)</td>
</tr>
<tr>
<td>cVAE-D</td>
<td>.760 (.010)</td>
<td>.776 (.015)</td>
<td>.760 (.010)</td>
<td>.778 (.015)</td>
<td>.758 (.010)</td>
</tr>
<tr>
<td>cVAE</td>
<td><strong>.776 (.009)</strong></td>
<td><strong>.794 (.015)</strong></td>
<td><strong>.773 (.010)</strong></td>
<td><strong>.790 (.017)</strong></td>
<td><strong>.774 (.009)</strong></td>
</tr>
<tr>
<td>Raw Data \ (d=562)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.758 (.009)</td>
<td>.780 (.015)</td>
</tr>
</tbody>
</table>
Case study: Disentanglement

Latent space: clustering (phenotyping) and disentanglement:

<table>
<thead>
<tr>
<th>Surgery Duration</th>
<th>Dimension 0 (Prediction)</th>
<th>Dimension 5</th>
<th>Other Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery type</td>
<td>Predicted surgery duration is represented in this dimension</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Case study: Disentanglement

- Latent space: clustering (phenotyping) and disentanglement:

<table>
<thead>
<tr>
<th>Surgery Duration</th>
<th>Dimension 0 (Prediction)</th>
<th>Dimension 5</th>
<th>Other Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Patients with similar surgery types are clustered together.

Surgery type information is represented in this dimension.

- Surgery type information is represented in this dimension.

- Surgery type information is represented in this dimension.
Case study: Disentanglement

- Latent space: clustering (phenotyping) and disentanglement

<table>
<thead>
<tr>
<th></th>
<th>Dimension 0 (Prediction)</th>
<th>Dimension 5</th>
<th>Other Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surgery</strong></td>
<td><img src="#" alt="Graph" /></td>
<td><img src="#" alt="Graph" /></td>
<td><img src="#" alt="Graph" /></td>
</tr>
<tr>
<td><strong>Duration</strong></td>
<td><img src="#" alt="Graph" /></td>
<td><img src="#" alt="Graph" /></td>
<td><img src="#" alt="Graph" /></td>
</tr>
<tr>
<td><strong>Surgery type</strong></td>
<td><img src="#" alt="Graph" /></td>
<td><img src="#" alt="Graph" /></td>
<td><img src="#" alt="Graph" /></td>
</tr>
</tbody>
</table>
Open problems

Model interpretation for OR-ICU handoff: ECHO project

- Given a prediction, how should clinicians trust/accept the model and interpret the predicted probability?
  - A predicted probability is not a probability
  - Is the risk high?
  - Why is the risk high?
  - How should I mitigate the risks?
Open problems

- Output calibration
  - Probabilities are not comparable between ML models
Open problems

Model interpretation for OR-ICU handoff: ECHO project

- Given a prediction, how should clinicians trust/accept the model and interpret the predicted probability?
  - A predicted probability is not a probability
  - Is the risk high?
  - Why is the risk high?
  - How should I mitigate the risks?
Summary

- **cVAE**: a representation learning framework for challenging clinical applications and others
  - Multi-modality
  - high dimensionality
  - performance critical
  - demanding interpretation

- Prediction-guided encoding
  - Improve predictive performance without a separate predictor

- Disentangled latent space
  - Interpretable phenotypes of patients
Related Publications
