Distilling Knowledge from Deep Networks with Applications to Healthcare Domain

Zhengping Che, Sanjay Purushotham, Yan Liu
University of Southern California
{zche, spurusho, yanliu.cs}@usc.edu

Motivations
- Electronic Healthcare Records (EHR) data is under exponential growth.
- New opportunities and urgent needs for discovery of meaningful data-driven representations and patterns in Computational Phenotyping research.
- Deep Learning models have shown superior performance for robust prediction in computational phenotyping tasks.
- Limited attempts on interpreting of features learned by deep learning models.
- Difficulties for clinicians in understanding and applying these models.
- Model interpretability is not only important but also necessary in healthcare.
- A good interpretable model is shown to result in faster adoptability among the clinical staff and is in better quality of patient care.
- Decision tree methods are widely employed in healthcare domain with easy interpretability but they do not achieve good performance.
- Question: How can we learn interpretable models from well-trained deep network models?
- Employ mimicking ideas suggested in recent deep learning papers, e.g., dark knowledge [2] and mimic learning [3].

Our Contributions
- Interpretable Mimic Learning - A simple yet effective knowledge distillation method
  - Mimic the performance of state-of-the-art deep learning models using well-known Gradient Boosting Trees (GBT).
  - Extensive experiments on several deep learning architectures
    - Include state-of-the-art deep networks: Stacked denoising autoencoders (SDA) and Long Short-Term Memory (LSTM).
  - Show Interpretable Mimic Learning models achieve comparable or even better performance than these deep learning models.
  - Interpretable features and decision rules, learned by our Interpretable Mimic Learning models, validated by expert clinicians.

Notations
- Assume each EHR data sample has static records with $P$ variables.
- By flattening the time series and concatenating static variables, we get an input vector $X \in \mathbb{R}^P$ for each sample, where $D = T + P$.
- We can also only focus on the temporal variables, with input $X_t = (x_{t1}, x_{t2}, \ldots, x_{tp})^T \in \mathbb{R}^P$, where $x_t \in \mathbb{R}^P$ represents the variables at time $t$.
- A binary label $y \in \{0, 1\}$ which represents the patient’s health state, e.g., mortality.

Methodology
- Feedforward Network (DNN) and Stacked Denoising Autoencoder (SDA) [4]
  - Transformation $X^{(l+1)}$ and reconstruction $Z^{(l)}$ of each layer $l$:
    \[ X^{(l+1)} = f_l(X^{(l)}) = (W_l^{(1)}X^{(l)} + b_l^{(1)}) \]
    \[ Z^{(l)} = g_l(W_l^{(2)}X^{(l)} + b_l^{(2)}) \]
- Long Short-Term Memory (LSTM) [5]
  - A popular recurrent neural network model for sequential data and tasks
  - Gradient Boosting Trees (GBT) [6]
    - An ensemble of weak learners (decision trees)
    - Find a linear combination of several functions $h(x)$ using gradient descent approaches to approximate the prediction function $F(x)$.
    - Final model with $M$ weak learners (stages): $F(x) = \sum_{l=1}^{M} \gamma_l h_l(x) + const$

Quantitative Results
- Dataset: Acute hypoxemic respiratory failure data of 398 child patients [7]
  - Two tasks:
    - MOR: Predict whether the patient dies within 60 days after admission.
    - VFD: Predict whether the patient survives and is on a ventilator for more than 14 days
  - Ventilator free Days is a surrogate outcome of morbidity and mortality.
- Classification Results: AUC(mean / std): Mean / Standard deviation of Area Under ROC
  - Other baseline methods: SVM, Support Vector Machine; LR, Logistic Regression; DT, Decision Tree

Interpretations
- Top useful features and corresponding importance scores
  - Top features with high importance scores for each task.

References