Minimizing HAI-Prevention Cost, with Probabilistic Guarantees

Elioth Sanabria and David D. Yao
IEOR Dept, Columbia Engineering

Healthcare-Associated Infection (HAI)

Healthcare associated infections (HAI) are estimated to cost US hospitals $9.8B per year. (Starting 2008, Medicare stopped reimbursing hospitals expenses due to HAI.) Costs for taking preventive measures are order-of-magnitude lower, but success rates are far from 100%; refer to the two tables below. We want to develop a machine learning (ML) classification scheme based on patient admission data only, as illustrated in the figure below.

<table>
<thead>
<tr>
<th>sources of infection (non-surgical)</th>
<th>cost per case</th>
<th>infection rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>catheter-associated urinary tract (CIU)</td>
<td>$996</td>
<td>1.15%</td>
</tr>
<tr>
<td>central line-associated bloodstream (CLAB)</td>
<td>$45,814</td>
<td>3.35%</td>
</tr>
<tr>
<td>clostridium difficile infection (CDIFF)</td>
<td>$11,205</td>
<td>0.37%</td>
</tr>
<tr>
<td>ventilator-associated pneumonia (VAP)</td>
<td>$40,144</td>
<td>0.25%</td>
</tr>
</tbody>
</table>

A Jointly Optimized Classification-Prediction Scheme

The model involves two interlaced optimization problems. At the top is a cost minimization problem that explicitly accounts for the asymmetry between the cost of infection and the cost of prevention. The infection probabilities used in the cost model are solutions to a cross-entropy (CE) minimization problem that fits data with a suitable ML algorithm (e.g., logit regression, random forest, deep neural network, etc). Here, the challenge is to deal with the intrinsic bias in data: infected cases are only around 1-2% of all patients. Our approach is to add a weighting (or "oversampling") coefficient to the CE objective and make it a decision variable too, in the same spirit as a Lagrangian multiplier.

Convergence and Rate of Convergence

\[ \phi^*: \text{optimal solution to the original problem;} \]
\[ \hat{\phi}_n^*: \text{optimal solution to the "data-driven" version;} \]
\[ EC(\phi^*), \hat{E}_n(\hat{\phi}_n^*): \text{corresponding objective values.} \]

Applying the Dvoretzky-Kiefer-Wolfowitz/Massart bound, we can derive

\[ \hat{E}_n(\hat{\phi}_n^*) \rightarrow EC(\phi^*) \ a.s.; \text{ and } P(\hat{E}_n(\hat{\phi}_n^*) - EC(\phi^*)) < \epsilon \leq 4 \exp \left( - \frac{n \epsilon^2}{2K^2} \right). \]

\[ C(\phi^*_n) \rightarrow EC(\phi^*) \ a.s.; \text{ and } P(\hat{C}(\phi^*_n) - EC(\phi^*)) > \epsilon \leq 4 \exp \left( - \frac{n \epsilon^2}{2K^2} \right). \]

Suppose \( \phi^*_n \) (for a given \( n \), sufficiently large) is applied to another data set of size \( N \), with the data being i.i.d. and following the same distribution as the original set, and denoting the corresponding cost as \( \hat{E}_N(\hat{\phi}_N^*) \). Then,

\[ \hat{E}_N(\hat{\phi}_N^*) \rightarrow \infty \ EC(\phi^*) \ a.s.; \text{ and } P(\hat{E}_N(\hat{\phi}_N^*) - EC(\phi^*)) > \epsilon \leq P \left( |Z| > \frac{\sqrt{3N}}{2\delta} \right), \]

where \( Z \): standard normal, \( \delta^2 := K_n(1-\pi) + K^2_n \pi \).

Cost Savings Achieved

As shown in the left figure below, logit regression achieves about 10% reduction, DNN does about 20%, from the base case (no ML). The right figure shows the confidence interval associated with the DNN performance.

Research Support and External Collaborations

Agency for Healthcare Research and Quality, AHRQ-R01-HS024915-01 (PI: Elaine Larson)
Columbia University Medical Center, School of Nursing