Enabling translational medicine using Bayesian Rule Learning with informative structure priors

Jeya Balasubramaniam

Intelligent Systems Program, University of Pittsburgh

March 23rd, 2018
Translational Medicine

A task of clinical importance is the prediction of diagnosis/prognosis of a disease.

Biomedicine has evolved into a data-intensive science with large amounts of data available from several sources.

Development of high-throughput ’omic’ technologies (eg., genomics, transcriptomics, proteomics, etc.) provides holistic view of a biological process.
Biomedical Data: Challenges

- An important challenge comes from the high-dimensionality of these datasets (large number of candidate variables, few instances).
- Data mining algorithms can get stuck in local optima.
- The algorithms can infer associations from spurious variables.
Some desirable properties of a learned predictive model in biomedicine include—

- **Good predictive performance**: Errors in biomedicine are expensive.
- **Model comprehensibility**: Ideally predictions are explained (not black-box) and offer testable hypothesis.
Domain knowledge sources (basic sciences)

Additional domain knowledge about the dataset can come from—

- A domain expert (e.g. a physician).
- Domain literature (e.g. PubMed).
- Domain knowledge-bases (e.g. Gene Ontology).
- Related datasets (e.g. GEO for gene expression datasets).
Enabling translational medicine

Problem statement
To enable translational medicine, we need intelligent agents\(^1\) (e.g. a data mining system) that can obtain background knowledge, from the basic sciences, and learn knowledge more effectively with this background knowledge to improve clinical practice.

The properties of the inferred model should include—
- Good predictive performance,
- Comprehensible.

---
Bayesian Rule Learning (BRL)

- BRL learns a constrained Bayesian network over a target variable of interest, from an input dataset, and infers a rule set from its conditional probability table.

- The BRL greedy best-first search is implemented as follows—
  - **Initialization**: Target variable node with no parent variables.
  - **Specialization**: One new candidate variable is added as parent to target variable of the best current model.
  - **Heuristic evaluation**: Bayesian score.
  - **Termination**: Maximum conjunct limit is reached (set to 8).

- Successfully applied to biomedical datasets (Floudas, 2013; Balasubramaian, 2014; Shi, 2014) and have shown to be better than state-of-the-art comprehensible models (Gopalakrishnan, 2010; Balasubramaian, 2017).
BRL: Constrained Bayesian network
BRL: Inferred rules (global structure)

1. IF (Gene A = UP) AND (Gene B = UP) THEN [0.89, 0.11] (50, 5)
2. IF (Gene A = UP) AND (Gene B = DOWN) THEN [0.50, 0.50] (4, 4)
3. IF (Gene A = DOWN) AND (Gene B = UP) THEN [0.34, 0.66] (15, 30)
4. IF (Gene A = DOWN) AND (Gene B = DOWN) THEN [0.34, 0.66] (10, 20)
The informed BRL (iBRL) framework
Bayesian Inference

The posterior probability of the candidate Bayesian network graph, $G$, given an observed dataset, $D$, is given by—

$$P(G|D) = \frac{P(G,D)}{P(D)}$$

$$P(G|D) \propto P(G,D)$$
Likelihood Score

\[ P(G, D) = P(G) \cdot P(D|G) \]

Likelihood function is computed using Bayesian Dirichlet equivalent uniform (BDeu) score by Heckerman et al. (1995)—

\[
P(G, D; \alpha) = P(G) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma\left(\frac{\alpha}{q_i}\right)}{\Gamma(N_{ij} + \frac{\alpha}{q_i})} \prod_{k=1}^{r_i} \frac{\Gamma\left(N_{ijk} + \frac{\alpha}{r_i q_i}\right)}{\Gamma\left(\frac{\alpha}{r_i q_i}\right)}
\]

where,

\( \alpha \), is the prior equivalent sample size.
(set to 1; Koller and Friedman (2009))
\[ N_{ij} = \sum_k N_{ijk} \]
Literature: Structure Priors

- Typically **uninformative priors** are used—

\[ P(G) \propto 1 \]

- Two notable works with **informative priors**—

1. Castelo and Siebes (2000) described a completed prior \((\hat{p}_{ij}, \overrightarrow{p}_{ij}, p^0_{ij})\).
   - Elicit prior probabilities for a subset of arcs.
   - Set remaining as a discrete uniform distribution.
2. Mukherjee and Speed (2008) proposed an informative prior using a log-linear combination of arbitrary features $f_i(\cdot)$ of candidate graph $G$—

$$P(G) \propto \exp \left( \lambda \sum_i w_i f_i(G) \right)$$

where,

- $w_i$, relative importance of the $i^{th}$ feature.
- $\lambda$, overall strength of the prior.
In iBRL, we calculate structure priors as follows—

\[ P(G) \propto \exp\left(\lambda \cdot (|E(G) \cap E_+| - |E(G) \cap E_-|)\right) \]

where,

- \( E(G) \), is the edge set in candidate graph \( G \).
- \( E_+ \), set of edges expected to be present.
- \( E_- \), set of edges expected to be absent.
Hyperparameters

- We set weights $w_i$ to 1.
- The $\lambda$ hyperparameter is related to Jeffrey’s scale used to describe the range of Bayes factors.
- In Bayesian approach to model selection, the Bayes factor, $K$, is used to evaluate strength of evidence for hypothesis $G_B$ over $G_A$ as follows—

$$K = \frac{P(D|G_B)}{P(D|G_A)}$$

- We test a range of $\lambda$ from 0 to 10.
We propose the iBRL framework, an intelligent agent, that can translate background knowledge from various sources, incorporate into the constrained Bayesian network model learning process in form of structure priors, to learn knowledge more effectively.

iBRL models are shown to—

- have good predictive performance,
- be comprehensible rule form.
Experiment

- Study the effect of prior background knowledge on model learning on a real-world dataset.
Dataset preparation

- We extract a lung cancer prognostic dataset from Gene Expression Omnibus (GSE19804).
- Data has 120 samples and 54,675 variables (probes) with 60 patients with lung cancer. It includes tissues from paired tumor and adjacent-normal cells.
- The data is normalized (Robust multi-array average) and probes are mapped to genes (using inter-quantile range filtering).
- The processed dataset has 120 samples and 16,383 variables (genes).
- In each cross-fold of our experiment, we discretize the dataset using efficient Bayesian discretization (parameter = 0.5). This acts as a feature selector and selects about 50% of the variables.
## Dataset clinical characteristics

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>Number of samples (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Women</td>
<td>60 (100%)</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Tumor type</td>
<td>Adenocarcinoma</td>
<td>56 (93%)</td>
</tr>
<tr>
<td></td>
<td>Bronchioloaveolar</td>
<td>3 (5%)</td>
</tr>
<tr>
<td></td>
<td>carcinoma</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Squamous</td>
<td>1 (2%)</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Stage</td>
<td>I + II</td>
<td>47 (78%)</td>
</tr>
<tr>
<td></td>
<td>III + IV</td>
<td>13 (22%)</td>
</tr>
<tr>
<td>Smoking history</td>
<td>Yes</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>60 (100%)</td>
</tr>
</tbody>
</table>
Semaphorins and their receptors have emerged as pivotal signals that are deregulated in cancer cells\textsuperscript{1}.

SEMA5A is a potential therapeutic target for lung cancer and a potential marker for NSCLC in non-smoking women\textsuperscript{2}.


Results: SEMA5A gene as prior
Epidermal growth factor receptor (EGFR) tyrosine kinase mutation is a good prognostic marker for both early and advanced-stage patients\textsuperscript{1–3}.

Advanced NSCLC patients are tested for EGFR mutations to develop personalized therapies.


Results: EGFR gene as prior

Effect of EGFR as prior on Average AUC across 5 runs of 10-fold cross-validation

AUC

Algorithm: iBRL

λ
Cooking oil fumes condensate (COFC) is implicated to be an important lung cancer risk factor in Chinese women.

Wu et al. (2008)\(^1\) performed an *in vitro* analysis to show that down-regulation of OGG1 gene leads to slow repair of DNA damage caused by COFC.

---

Results: OGG1 gene as prior
Results: Comparison to baseline models

![Average AUC over 5 runs of 10-fold cross-validation (with standard error)](image)
Benefits of iBRL

- Allows for incorporation of prior domain knowledge from a wide variety of sources.
  - Some domain knowledge may help escape local optima.
- Can account for the uncertainty in the validity of the prior knowledge (tunable using hyperparameters $w_i$ and $\lambda$).
- Novel biomarker discovery (penalize known knowledge).
Future Work

- Combine multiple sources of prior knowledge \((f_i(\cdot))\) and study their effect on model learning.
- Estimate values of hyperparameters \((w_i, \lambda)\) in a more informed way.
- Iterative framework where we can focus on subpopulations that are poorly predicted by the model.
Conclusion

- We proposed an intelligent agent, the iBRL, which provides a flexible framework to incorporate prior domain knowledge into the model learning process.
- We demonstrate the usefulness of iBRL in describing suspected prognostic markers for the development of lung cancer.