Diabetes Prediction with Incomplete Patient Data

Hao Yi Ong, Dennis Wang, Xiao Song Mu
CS 221 Artificial Intelligence: Principles and Techniques Class Project

Introduction

- Given a set of electronic health records, we want to have a smart predictor that prompts high-risk patients to obtain Type II Diabetes testing

- Original Kaggle challenge:
  - Patients all have a standard database and a full medical record
  - I.e., exact same tests taken, same variables recorded, etc.
- In practice, however:
  - Not everyone has taken the same tests and gotten regular checkups
  - Database inconsistencies or errors in inputting data may exist
- Predictor must be able to accurately classify based on incomplete or erroneous medical records to be useful

Diabetes Prediction

- Given
  - *Training set* containing standardized patient medical records
  - *Testing set* containing patient medical records with missing informati and unknown erroneously recorded data
- Output
  - *Bayesian network structure* encoding the conditional dependencies between medical record variables
  - *Bayesian network parameters* encoding the conditional probabilities for each variable
  - *Probabilistic inference* on the learned Bayesian network (BN) for classification
- To minimize the error rate, including false positives and false negatives, on classifying whether a patient has Type II Diabetes

Evaluation Criteria

- *Baseline* of logistic regression
  - Feature vector with basic data including height, weight, body mass index...
  - Cross validation with 10%-hold-out on the training data
  - Obtained a false positive rate of 0.7% and false negative rate of 15.3% for a total error rate of 16% (84% accuracy)
- *Oracle*
  - Ideal oracle would be experienced physicians who have correctly advised patients to test for diabetes given their medical history (impractical)
  - Use as surrogate measures the test accuracies of established diabetes tests vetted by the US Department of Health and Human Services
  - Specifically, HBA1c, FPG, and OGTT, which have 85–95% accuracy

Structure Learning

Structure learning as a search problem

- **Feature selection** We manually selected 27 discretized features according to ICD9 disease groupings to limit the search space.
- **Bayesian score** To evaluate the BN, we use the Bayesian score as the scoring function, which optimally balances the complexity of the BN structure with the available data (Koller and Friedman, 2009). The function is complicated and we do not include it here for brevity's sake.
- **NP-hard** Searching for the globally optimal structure is equivalent to searching over the space of all possible directed acyclic graphs, whose cardinality grows superexponentially with the number of nodes.
- **Tabu search** To find the optimal structure, we use a hill-climbing algorithm that maximizes the "fitness" of the BN based on the Bayesian score.
  By maintaining a tabu list of recent operators we applied (e.g., adding an edge to the existing BN) and not considering operators that reverse the effect of recently applied operators, the heuristic avoids getting stuck at local optima.

given $\mathcal{X}^{\text{fixed}}$, $\mathcal{F}$, $\hat{\mathcal{D}}$

Generate set of node coordinates x^0 from $\hat{\mathcal{D}}$, set $x := x^0$

repeat

1. Given x, obtain a and Θ_1 as the solution to and objective of (**??**)
2. Given a, obtain y and Θ_2 as the solution to and objective of (**??**), set $x := x + y$
3. **break if** Θ_1 and Θ_2 converge

return a, x

Parameter Learning

Approximate Probabilistic Inference

Conclusion

Acknowledgments

We thank Professor Liang and the instructor team, as well as fellow classmates for their help on our project.