A Probabilistic Approach to Extract Qualitative Knowledge for Early Prediction of Gestational Diabetes

Athresh Karanam*[1], Alexander L. Hayes*[2], Harsha Kokel[1], David M. Haas[2], Predrag Radivojac[3], and Sriraam Natarajan[1]

[1] The University of Texas at Dallas, USA
[2] Indiana University Bloomington, USA
[3] Northeastern University, USA

Abstract. Qualitative influence statements are often provided a priori to guide learning; we answer a challenging reverse task and automatically extract them from a learned probabilistic model. We apply our Qualitative Knowledge Extraction method toward early prediction of gestational diabetes on clinical study data. Our empirical results demonstrate that the extracted rules are both interpretable and valid.

1 Introduction

The nuMoM2b (Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be) study [3] aims to identify early warning signs of adverse pregnancy outcomes, design interventions, and assist with decision-making. Since 2010, eight research sites in the United States followed up with women throughout their pregnancies—collecting routine clinical information, exercise data, and food they ate. Using this data, we consider learning to explain the relationship between gestational diabetes mellitus (GDM) and some common risk factors.

A common way to employ knowledge in machine learning and AI is via the use of qualitative relationships that express how changes in a (subset of) feature(s)/risk factor(s) affect the target. These rules were mainly used as "inductive bias" apriori to learning since they are both intuitive and natural in many domains. We address the challenging "reverse task". Can we extract these rules from data? To this effect, in the context of nuMoM2b, we propose a two step process. First we learn a joint probability distribution over all the variables including the target (GDM). In the second step, the constraints are extracted by reasoning over this joint probability distribution. We demonstrate in our experiments that such an approach yields rules that are both intuitive and valid (as validated by our clinical expert Dr. David Haas). We first explain these constraints before outlining our approach and presenting our learned rules.

2 Extracting Qualitative Influences

A qualitative influence (QI) statement outlines how a change in one or more factor(s) would influence another factor [8]. We focus on two types of QI: *Mono-*

* Equal contribution

tonicity and Synergy [1, 5, 9]. *Monotonicity* represents a direct relationship between two variables: "As BMI increases, neck circumference increases" indicates that the probability of greater neck circumference increases with increase in BMI. Specifically, a *monotonic influence* (MI) of variable X on variable Y, denoted by $X \overset{M+}{\prec} Y$ (or its inverse $X \overset{M-}{\prec} Y$), indicates that higher values of X stochastically result in higher (or lower) values of Y. *Synergy* represents interactions among influences. Two variables synergistically influence a third if their joint influence is greater than their separate, statistically independent influences. Synergy can capture influences like "Increase in BMI increases the risk of high blood pressure in patients with family history of hypertension more than patients without family history." Formally, a *synergistic influence* (SI) of two variables A and B on variable Y, denoted by $A, B \overset{S+}{\prec} Y$, indicates that increasing the value of A has greater effect on Y for higher value of B than the lower value of B. Both A and B should necessarily have same monotonic relationship with Y.[4] Similarly, a *sub-synergistic influence* (sub-SI), denoted by $A, B \overset{S-}{\prec} Y$, indicates that while A and B have increasing monotonic influence on Y, the joint influence is lesser than their separate, statistically independent influence.

2.1 Proposed Approach

Given: A data set $\mathcal{D}$ consisting of examples in the form of risk factors $\mathbf{X}$ and binary target Y (in this case: GDM).
To Do: Learn a set of QIs that explain the effect of $\mathbf{X}$ on Y.

We use X_a to denote the a^{th} variable in the feature set $\mathbf{X}$. x_a^i denotes a particular value of variable X_a and $|X_a|$ denotes the number of discrete values X_a takes. We assume that the joint distribution (P) over the set of random variables $\mathbf{X}$ is known (we learn this joint distribution in our empirical evaluation using a causal learning algorithm). For brevity, we restrict the description of our method to extracting positive MIs and SIs, $\prec^{M+}$ and $\prec^{S+}$. The *degree of monotonic influence*, δ_a, of $X_a \in \boldsymbol{X}$ on Y is defined as

$$\delta_a = I_{(C_a>0)} \cdot \sum_j \sum_{j'>j} \sum_k \frac{P(Y \leq k | X_a = x_a^j) - P(Y \leq k | X_a = x_a^{j'})}{|X_a|} \quad (1)$$

where,

$$C_a = \prod_j \prod_{j'>j} \prod_k \max(P(Y \leq k | X_a = x_a^j) - P(Y \leq k | X_a = x_a^{j'}) + \epsilon_m, 0) \quad (2)$$

For monotonicity to hold, we require $P(Y \leq k | X_a = x_a^j) + \epsilon_m \geq P(Y \leq k | X_a = x_a^{j'})$ for all pairs of configurations of $X_a, (j, j')$ with $j' > j$ at any given threshold value k. Here the monotonic slack ϵ_m allows violating a constraint within a chosen margin. The degree of MI, δ_a, in Equation 1 measures the cumulative difference in the probability that the target variable Y is less than a threshold k given X_a at two different values x_a^j and $x_a^{j'}$.

We extend the concept of degree of MI to SI by conditioning on a pair of variables instead of a single variable. First, consider the difference in the effect

[4] Without loss of generality, assume the variables in synergistic relation have monotonically increasing impact.

of changing X_a from x_a^i to $x_a^{i'}$ on Y under the context of two different values of X_b (x_b^j and $x_b^{j'}$). We define this as

$$\phi_{a,b}^{i,i',j,j'} = \sum_k P(Y \leq k|X_a = x_a^i, X_b = x_b^j) - P(Y \leq k|X_a = x_a^{i'}, X_b = x_b^j) - P(Y \leq k|X_a = x_a^i, X_b = x_b^{j'}) + P(Y \leq k|X_a = x_a^{i'}, X_b = x_b^{j'})$$

For synergy to hold, we require $\phi_{a,b}^{i,i',j,j'} + \epsilon_s$ to be non-negative for all $i' > i$ and $j' > j$. Where ϵ_s is the synergistic slack. We define the *degree of synergistic influence*, $\delta_{a,b}$, of variables $X_a \in \boldsymbol{X}$ and $X_b \in \boldsymbol{X}$ on $Y \in \boldsymbol{X}$ as the cumulative difference in degrees of context specific influence of X_a on Y in the context of X_b. It is given by

$$\delta_{a,b} = I_{(C_{a,b}>0)} \cdot \sum_i \sum_{i'>i} \sum_j \sum_{j'>j} \frac{\phi_{a,b}^{i,i',j,j'}}{|X_a| \cdot |X_b|} \tag{3}$$

where,

$$C_{a,b} = \prod_i \prod_{i'>i} \prod_j \prod_{j'>j} \max(\phi_{a,b}^{i,i',j,j'} + \epsilon_s, 0) \tag{4}$$

We employ both definitions to learn QIs in Algorithm 1, Qualitative Knowledge Extraction (QuaKE). The algorithm assumes the existence of a joint distribution [6] over ordinal features, which we learn using a causal probabilistic learning algorithm (PC) [7, 2]. We chose PC algorithm to verify our hypothesis that the use of a causal model will yield causally interpretable qualitative relationships. We calculate the degree of MI of every variable $X_a \in \boldsymbol{X}$ on Y and SI of every pair of variables $X_a, X_b \in \boldsymbol{X}$ on Y. The MI rules $X_{a\prec}^{M+}Y$ are extracted if their corresponding degree of MI δ_a are above a pre-defined threshold T_m. Similarly, the synergistic rules $X_a, X_{b\prec}^{S+}Y$ are extracted if their corresponding degree of SI $\delta_{a,b}$ are above a pre-defined threshold T_s.

Algorithm 1: QuaKE

```
input : P, Y, X, ε_m, ε_s, T_m, T_s
output: Rules R
initialize: R ← ∅
for a ← 0 to (|X| − 1) do
    compute δ_a using Eq. 1
    if δ_a ≥ T_m then
        R ← (X_{a≺}^{M+} Y) ∪ R
    for b ← a + 1 to (|X| − 1) do
        compute δ_{a,b} using Eq. 3
        if δ_{a,b} ≥ T_s then
            R ← (X_a, X_{b≺}^{S+} Y) ∪ R
// Decreasing cases
return R
```

3 Learning qualitative influences for GDM modeling

The nuMoM2b study tracked pregnancies of 10,037 women near 8 sites in the United States. We excluded 817 cases where women were already diagnosed with diabetes; and we evaluate our proposed method for extracting QIs using 8 features[5] of the remaining 9,220 women. 7 features had inherent ordering of categories whereas *Race* had no obvious ordering. We use an ordering based on previous studies [4] on the effect of *Race* on *GDM*.

[5] Refer to the supplementary material for details on the data and features: https://starling.utdallas.edu/papers/QuAKE/

Rule	Prior Knowledge	QuaKE	Data Alone
$BMI \prec^{M+} GDM$	✓	✓	✓
$Age \prec^{M+} GDM$	✓	✓	✓
$Race \prec^{M+} GDM$	✓	✓	✗
$Education \prec^{M+} GDM$	✓	✓	✗
$Gravidity \prec^{M+} GDM$	✓	✓	✗
$Smoked3months \prec^{M+} GDM$	✓	✗	✗
$SmokedEver \prec^{M+} GDM$	✓	✗	✗
$Age, BMI \prec^{S+} GDM$	✓	✓	✓
$Age, Smoked3months \prec^{S+} GDM$	✓	✓	✓
$BMI, SmokedEver \prec^{S+} GDM$	✓	✓	✓
$Education, Smoked3months \prec^{S+} GDM$	?	✓	✓
$BMI, Gravidity \prec^{S+} GDM$	✓	✓	✗
$BMI, Smoked3months \prec^{S+} GDM$	✓	✗	✓
$Age, SmokedEver \prec^{S+} GDM$	✓	✗	✗
$BMI, Education \prec^{S+} GDM$	✗	✓	✓
$Education, SmokedEver \prec^{S+} GDM$	?	✗	✗
$Age, Education \prec^{S-} GDM$	✓	✓	✓
$BMI, Smoked3months \prec^{S-} GDM$	✗	✓	✗
$Age, SmokedEver \prec^{S-} GDM$	✗	✗	✓
$BMI, Gravidity \prec^{S-} GDM$	✗	✗	✓
$Gravidity, SmokedEver \prec^{S-} GDM$	✗	✗	✓
$Education, SmokedEver \prec^{S-} GDM$	?	✗	✓
$Age, Gravidity \prec^{S-} GDM$	✓	✗	✗

Table 1. Comparision of QI from prior knowledge (PK), QuaKE and Data Alone. ✓/✗ represents that this relationship does/not exist respectively while **?** represents unknown influence. The three groups of rows show: (1) MI, (2) SI, and (3) sub-SI. Colors highlight rules recovered by QuaKE and show (a.) coherent with the PK and baseline (b.) contradicting the baseline (c.) coherent with baseline but contradicts the PK.

We pose and answer the following questions: **(Q1)** Does QuaKE extract high-quality rules that align with background knowledge in this domain? **(Q2)** Does QuaKE help uncover QI statements in cases where prior knowledge is uncertain?

We compare learned rules with those from our clinical expert, *Dr. Haas*. W.r.t GDM, these could either be increasing, decreasing, no effect, or unknown. Since Algorithm 1 assumes a complete joint distribution P is available, we consider two factorizations of P. The first learns a causal model [2] and the other (baseline) estimates the probabilities directly from data. Alternative baselines might have included rules extracted from decision trees, rule mining, or Bayesian rule learning—but each induce conjunctive rules of the form $(x_1 \wedge x_2 \wedge ... \wedge x_n) \implies y$, making their exact connection to the QI statements tenuous.

All rules are presented in Table 1. The "Prior" knowledge refers to the rules provided by our expert. We compare these to the rules extracted by QuaKE and baseline (Data Alone). QuaKE's precision compared to expert advice is 0.923 ± 0; whereas the precision of our unstructured baseline is 0.636 ± 0. Precision of each

method was consistent across five stratified cross validation folds. This affirms **Q1**: QuaKE can extract high-quality rules aligning with prior knowledge.

Since we have formalized degree of the QIs in Equations 1 and 3, we can analyze rules that were highly uncertain according to the prior knowledge. Two of the synergistic relations involving smoking and education had an unknown effect with relation to GDM. $Education, Smoked3months \overset{S^+}{\prec} GDM$ was a high-confidence rule extracted by QuaKE and the baseline. We speculate that this could be either due to the high correlation between *Education* and *Age*, or related to an unobserved relationship between education and socioeconomic status. Note that both these results are especially interesting since we found only a weak monotonic relationship between smoking and GDM more generally. We use this to answer **Q2**: our approach can identify potentially interesting cases where prior knowledge is uncertain.

Discussion and Conclusion: We considered the problem of learning interpretable and explainable qualitative rules for modeling GDM. To this effect, we learned a causal (probabilistic) model and recovered the knowledge by applying the rules. Our results indicate that most of our rules are in line with the prior knowledge of our expert and some interesting influence relationships appear that are worth investigating. Incorporating richer domain knowledge, automatically refining the rules, identifying broader relationships and scaling to larger feature sets are interesting future research directions.

Acknowledgements: We gratefully acknowledge the support of 1R01HD101246 from NICHD and Precision Health Initiative of Indiana University. Thanks to Rashika Ramola, Rafael Guerrero for data processing and discussions.

References

1. Altendorf, E.E., Restificar, A.C., Dietterich, T.G.: Learning from sparse data by exploiting monotonicity constraints. In: UAI. pp. 18–26 (2005)
2. Colombo, D., Maathuis, M.H.: Order-independent constraint-based causal structure learning. Journal of Machine Learning Research **15**(1), 3741–3782 (2014)
3. Haas, D.M., Parker, C.B., et al.: A description of the methods of the nulliparous pregnancy outcomes study: monitoring mothers-to-be (nuMoM2b). American journal of obstetrics and gynecology **212**(4), 539.e1–539.e24 (2015)
4. Hedderson, M.M., Darbinian, J.A., Ferrara, A.: Disparities in the risk of gestational diabetes by race-ethnicity and country of birth. Paediatric and Perinatal Epidemiology **24**(5), 441–448 (2010)
5. Kokel, H., Odom, P., Yang, S., Natarajan, S.: A unified framework for knowledge intensive gradient boosting: Leveraging human experts for noisy sparse domains. In: AAAI. vol. 34, pp. 4460–4468 (2020)
6. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)
7. Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review **9**(1), 62–72 (1991)
8. Wellman, M.P.: Fundamental concepts of qualitative probabilistic networks. Artificial Intelligence **44**(3), 257–303 (1990)
9. Yang, S., Natarajan, S.: Knowledge intensive learning: Combining qualitative constraints with causal independence for parameter learning in probabilistic models. In: ECML-PKDD. vol. 8189, pp. 580–595 (2013)