Моделирование неточности ответов в байесовской сети доверия для оценивания социально-значимого поведения респондентов

Акульшина О.Д, студент СПбГУ, стажёр-исследователь СПИИРАН, 9.svarog@gmail.com[1]

Аннотация

Рассмотрено влияние неточности ответов респондентов на качество предсказаний модели, основанной на байесовской сети доверия.

Введение

Во многих задачах психологии, социологии, эпидемиологии возникает необходимость измерения параметров социально-значимого поведения индивида [1, 8]. Одним из подобных примеров является оценка риска получения или передачи ВИЧ-инфекции на основе данных о последних эпизодах рискованного поведения. Но непосредственное измерение числа эпизодов социально-значимого поведения часто невозможно. Для предсказания числа таких эпизодов в работах [2, 3] предложена модель на основе байесовской сети доверия, которая позволяет оценить интенсивность поведения респондентов по небольшому числу эпизодов. Однако оценки интенсивности, построенные согласно такому подходу могут быть искаженны, по причине того, что ответы, полученные от респондентов, не всегда соответствуют реальности: респондент может не помнить точного ответа на вопрос или намеренно дать неверный ответ.

Таким образом, целью данной работы является изучение влияния искажённых данных, полученных от респондентов, на качество предсказания модели.

Модель на основе байесовской сети доверия

Исходными данными в существующей модели, описанной в работах [2, 3], являются сведения о интервалах между тремя последними

[1]Работы выполнялись в рамках проекта по государственному заданию СПИИРАН № 0073-2014-0002.

эпизодами рискованного поведения (t_{01}, t_{12}, t_{23}) и сведения о минимальном и максимальном интервале ($t_{\min}$, $t_{\max}$) между эпизодами за исследуемый период времени. Кроме того, модель содержит оцениваемую величину, соответствующую интенсивности поведения (λ) и число эпизодов, произошедших за исследуемый промежуток времени (n).

Для построения байесовской сети значения непрерывных величин (λ, t_{01}, t_{12}, t_{23}, $t_{\min}$, $t_{\max}$) разбиваются на дискретные промежутки. В данной работе рассматривается следующая дискретизация:

λ : $\lambda^{(1)} = [0, 0.01)$, $\lambda^{(2)} = [0.01, 0.03)$, $\lambda^{(3)} = [0.03, 0.05)$, $\lambda^{(4)} = [0.05, 0.1)$, $\lambda^{(5)} = [0.1, 0.2)$, $\lambda^{(6)} = [0.2, 0.3)$, $\lambda^{(7)} = [0.3, 0.5)$, $\lambda^{(8)} = [0.5, 0.7)$, $\lambda^{(9)} = [0.7, 1)$, $\lambda^{(10)} = [1, \infty)$

t_{01}, t_{12}, t_{23}, $t_{\min}$, $t_{\max}$: $t^{(1)} = [0, 0.1)$, $t^{(2)} = [0.1, 1)$, $t^{(3)} = [1, 7)$, $t^{(4)} = [7, 30)$, $t^{(5)} = [30, 180)$, $t^{(6)} = [180, \infty)$

Случайная величина n представляет собой натуральный логарифм числа эпизодов за рассматриваемый период времени, и разбивается на 10 равных дискретных промежутков, в зависимости от полученных сведений.

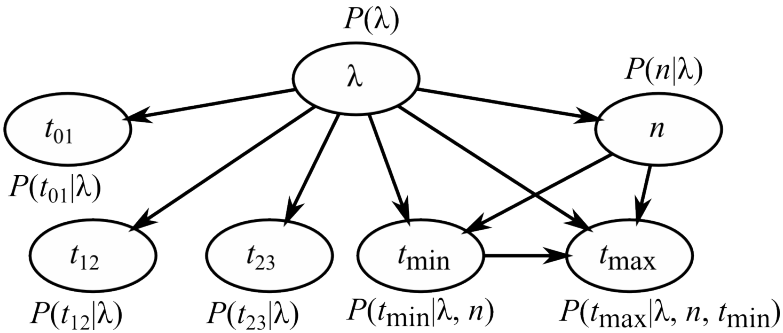

Рис. 1: Структура БСД для моделирования социально-значимого поведения

Таким образом, данную модель (рис. 1) можно представить в виде байесовской сети доверия с экспертно заданной структурой, представленной графом $G = (V, L)$, где $V = \{\lambda, n, t_{01}, t_{12}, t_{23}, t_{\min}, t_{\max}\}$ — множество вершин, а $L = \{(u, v) : u, v \in V\}$ — множество направленных дуг, показывающих связи между элементами.

Описание тестовых данных

Для тестирования качества предсказания описанной выше модели используются сгенерированные данные об эписодах рискованного по-

ведения, соответствующие теоретическим предположениям

Генерация неискажённых данных

Для генерации данных используется программа, представленная в работе [4], которая генерирует эпизоды поведения в соответствии с теоретическими предположениями модели.

Сначала генерируются значения интенсивности, соответствующие значениям случайной величины, распределенной по гамма-распределению с параметрами $k = 1.1$, $\Theta = 0.3$. С одной стороны, большая часть значений меньше 0.5, что соотносится со многими примерами реального поведения, с другой — в данных есть значения для всех интервалов, на которые разбито значение λ при дискретизации.

Далее для каждого значения интенсивности генерируются респонденты — последовательности точек, расстояния между которыми подчиняются экспоненциальному распределению с соответствующим значением интенсивности. Из каждой такой последовательности выделяются исходные данные для оценки: длины интервалов между тремя последними точками, минимальный и максимальный интервал за промежуток длиной 365 дней, удаляются последовательности, у которых нет хотя бы двух точек за этот промежуток.

Таким образом формируется набор переменных rate, n, le1, le2, le3, min, max, соответствующий случайным элементам модели λ, n, t_{01}, t_{12}, t_{23}, $t_{\min}$, $t_{\max}$.

Для генерации данных для обучения генерируется 300 значений интенсивности и 20 респондентов для каждого значения интенсивности.

Для генерации данных для тестирования генерируется 50 значений интенсивности и 15 респондентов для каждого значения интенсивности.

Генерация искажённых данных

Для создания искажённых данных из уже сгенерированных используются функции, которые принимают набор, состоящий из переменных le1, le2, le3, min, max, описывающих интервалы между тремя последними эпизодами рискованного поведения и максимальный и минимальный интервал между такими эпизодами за указанный период времени и добавляет «шум» к этим значениям.

В данной работе рассмотрены следующие подходы к генерации искажённых данных:

1. Генерируется одно значение случайной величины, равномерно распределённой на отрезке $[0.75 * \text{le}, 1.25 * \text{le}]$, где le — длина, рассматриваемого в данный момент промежутка между двумя последними эпизодами поведения. Таким образом, получается новое «зашумлённое» значение, которое отклоняется от исходного значения не более, чем на четверть.

2. Генерируется одно значение нормально распределённой случайной величины с параметрами $\mu = 0$, $\sigma = 0.167 * \text{le}$, где le — длина, рассматриваемого в данный момент промежутка. Отметим, что плотность вероятности случайной величины, описанной выше, симметрична отночительно 0, и большинство её значений расположены в промежутке $[-0.5*\text{le}, 0.5*\text{le}]$. То есть большинство искажённых значений длины рассматриваемого исходного интервала будет отличаться не более, чем на половину.

3. Генерируется одно значение случайной величины с бета-распределением с параметрами $\alpha = 1$, $\beta = 3$, причём плотность вероятности такой случайной величины убывает на промежутке $[0, 1]$. Далее полученное значение случайной величины умножается на длину расматриваемого промежутка между двумя эпизодами социально-значимого поведения и случайно выбранный коэффициент 1 или -1 и прибавляется к рассматриваемому промежутку. Таким образом, полученное искажённое значение рассматриваемого промежутка иежду двумя эпизодами будет отличаться от исходного не более чем в два раза.

Такой подход к генерации «шума» имеет смысл, так как он отражает реальную жизнь: чем раньше от данного момента произошло событие, тем сложнее вспомнить его подробности; чем раньше от момента интервьюирования произошёл эпизод социально-значимого поведения, тем сложнее респонденту будет вспомнить точное время и тем больше возможно расхождение между истиной и ответом.

Сравнение предсказаний модели для различных тестовых данных

Для исследования влияния неточности респондентов на качество предсказания модели происходит автоматическое обучение параметров модели с помощью сгенерированного обучающего набора данных,

а затем на модели проводится предсказания интенсивности социально-значимого поведения на неискажённых и искажённых данных и сравнение результатов предсказания. Отметим, что работа выполняется на языке **R** [6], для обучения модели, основанной на байесовской сети доверия, и предсказаний используется пакет **bnlearn** [7].

Далее оценим формальные показатели качества предсказания модели, такие как accuracy, precision, recall, а также средняя по классам точность. Сводная таблица качеств предсказания представлена в таблице 1.

Таблица 1: Качество предсказания модели на различных наборах данных

	accuracy	precision	recall	accuracy (средняя)
Исходные данные	0.705	0.450	0.469	0.905
Данные с равномерным «шумом»	0.656	0.340	0.374	0.901
Данные с нормальным «шумом»	0.678	0.391	0.416	0.904
Данные с «шумом» с бета-распределением	0.667	0.380	0.397	0.897

Как можно увидеть из таблицы 1, качество предсказания моделей, обученных на «чистых» на «зашумлённых» данных отличаются друг от друга незначительно.

Заключение

Проведённое исследование показывает хорошую согласованность между предсказаниями на искажённых и неискажённых данных. Другими словами, если с точки зрения практической задачи снижение точности предсказания приблизительно на 5% не является критичным, то можно не предпринимать дополнительных действия для учета возможной зашумленности в данных, а использовать первоначальную модель. Однако, дополнительного исследования требуют случаи, когда шум может быть смешанным и разным у разных респондентов.

Литература

[1] Тулупьева Т.В., Пащенко А.Е., Тулупьев А.Л., Красносельских Т.В., Казакова О.С. Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей. СПб.: Наука, 2008. 140 с.

[2] Суворова А.В., Тулупьев А.Л. Структурный синтез байесовской сети доверия по пуассоновской модели поведения // XV национальная конференция по искусственному интеллекту с международным участием КИИ-2016 (3–7 октября 2016 г., г. Смоленск, Россия). Труды конференции. В 3-х томах., Смоленск: Универсум, 2016, Т. 3. С. 139–147

[3] Суворова А.В., Тулупьев А.Л., Сироткин А.В. Байесовские сети доверия в задачах оценивания интенсивности рискованного поведения // Нечеткие системы и мягкие вычисления. 2014. Т. 9, N2. С. 115–129.

[4] Суворова А.В. Гибридные оценки параметров социально-значимого поведения по сверхмалой неполной совокупности наблюдений // Труды СПИИРАН. 2013. Вып. 1(24). С. 116–134.

[5] Тулупьев А.Л., Николенко С.И., Сироткин А.В. Байесовские сети: логико-вероятностный подход. Спб.: Наука, 2006. 607 с.

[6] RStudio: Integrated development environment for R (Version 0.98.1060) [Computer software]. Boston, MA. Available from http://www.rstudio.org

[7] bnlearn: Bayesian network structure learning, parameter learning and inference. (Version 4.2) [Computer software]. Marco Scutari, Robert Ness. Available from http://www.bnlearn.com

[8] Varghese B., Maher J.E., Peterman T.A., Branson B.M., Steketee R.W. Reducing the risk of sexual HIV transmission: quantifying the per-act risk for HIV on the basis of choice of partner, sex act, and condom use // Sexually transmitted diseases. 2002. Vol. 29(1). P. 38–43.