Моделирование рискованного поведения на основе байесовской сети доверия: индивидуальные и групповые оценки интенсивности [1]

Суворова А.В., с. н. с. лаборатории ТиМПИ СПИИРАН,
suvalv@gmail.com

Аннотация

Предложена модель на основе байесовской сети доверия для оценивания характеристик рискованного поведения по данным об ограниченном числе эпизодов такого поведения. Рассмотрены примеры вычисления и интерпретации как для случая оценивания интенсивности поведения одного респондента, так и для получения оценки для целой группы. Проведено тестирование модели на сгенерированных данных.

Введение

В рамках многих социологических, психологических, маркетинговых исследований возникает необходимость моделирования поведения респондентов для последующего оценивания параметров такого поведения [1, 4, 9]. В работах [2, 3] предложена модель на основе байесовской сети доверия для косвенного оценивания интенсивности поведения респондентов по сведениям о небольшом числе эпизодов такого поведения, однако для дальнейшего практического ее использования необходимо оценить качество модели. Отметим, что во многих случаях получение исходных данных связано с рядом проблем, т.к. чаще всего данные об эпизодах получены из результатов опросов или интервью, что не позволяет собирать большие объемы данных в короткие сроки. Как следствие, усложняется процесс тестирования разработанной модели.

Цель данной работы — апробировать на генерируемых тестовых данных предложенную байесовскую сеть доверия для моделирования рискованного поведения респондентов на основе данных об эпизодах их поведения.

[1] Статья содержит материалы исследований, частично поддержанных грантами РФФИ 16-31-60063, 16-31-00373.

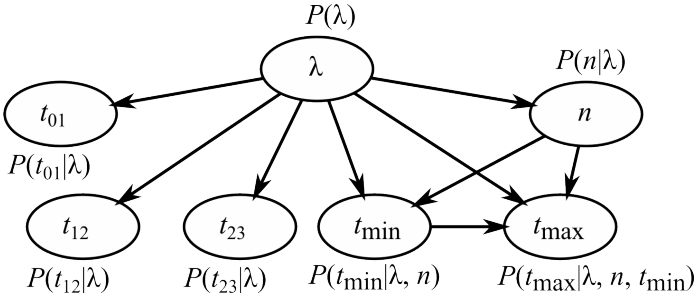

Рис. 1: Вероятностная графическая модель рискованного поведения.

Модель на основе байесовской сети доверия

Модель на основе байесовской сети доверия для решения задачи оценивания интенсивности рискованного поведения по сведениям об эпизодах такого поведения предложена в работах [2, 3]. Граф $G(V, L)$, где $V = \{t_{01}, t_{12}, t_{23}, t_{\min}, t_{\max}, \lambda, n\}$, $L = \{(u, v) : u, v \in V\}$, представляет структуру модели (рис. 1), где λ — случайная величина, характеризующая интенсивность поведения; t_{ij} — случайная величина, характеризующая длину интервала между i-ым и j-ым с конца эпизодами (0 обозначает момент интервью); $t_{\min}$ и $t_{\max}$ — случайные величины, характеризующие длины минимального и максимального интервалов соответственно; n — случайная величина, характеризующая число эпизодов за заданный промежуток времени.

Моделью поведения выступает пуассоновский случайный процесс, то есть длины интервалов между эпизодами распределены экспоненциально. Момент интервью и эпизоды поведения являются независимыми величинами (предполагается, что жизненные циклы респондента и интервьюера не связаны).

Тензоры условной вероятности, характеризующие переходы между узлами сети, где $\mathbf{P} = \left\{P\left(t_{j,j+1}\,|\lambda\right), P\left(t_{01}\,|\lambda\right), P\left(t_{\min}\,|n, \lambda\right), P\left(t_{\max}\,|n, \lambda, t_{\min}\right), P\left(n\,|\lambda\right), P\left(\lambda\right)\right\}$ определяются следующим образом ($l_s = 1, ..., k_s$, где k_s — число дизъюнктных промежутков при дискретизации случайных величин; $s = 0, ..., 4$; $j = 1, ..., 2$; $i = 1, ..., m$, где m — число дизъюнктных промежутков при дискретизации величины λ) [2]:

$$p\left(t_{j,j+1}^{(l_j)} \middle| \lambda^{(i)}\right) = e^{-a\lambda^{(i)}} - e^{-b\lambda^{(i)}}, j = 0, 1, 2, t_{j,j+1}^{(l_j)} = [a; b)\,;$$

$$p\left(t_{\min}^{(l_3)}\left|n,\lambda^{(i)}\right.\right)=e^{-an\lambda^{(i)}}-e^{-bn\lambda^{(i)}},t_{\min}^{(l_3)}=[a;b)\,;$$

$$p\left(n\left|\lambda^{(i)}\right.\right)=\frac{\left(\lambda^{(i)}T\right)^n}{n!}e^{-\lambda^{(i)}T};$$

$$p\left(t_{\max}^{(l_4)}\left|n,\lambda^{(i)},t_{\min}^{(l_3)}\right.\right)=e^{(n-1)\lambda^{(i)}t_{\min}^{(l_3)}}\times$$

$$\times\left(\left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}}-e^{-\lambda^{(i)}b}\right)^{n-1}-\left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}}-e^{-\lambda^{(i)}a}\right)^{n-1}\right),\ t_{\max}^{(l_4)}=[a;b)\,.$$

Во всех примерах, рассмотренных в данной работе, используется дискретизация вида: для случайной величины, соответствующей интенсивности поведения $\lambda^{(1)}=[0;0,01)$, $\lambda^{(2)}=[0,01;0,03)$, $\lambda^{(3)}=[0,03;0,05)$, $\lambda^{(4)}=[0,05;0,1)$, $\lambda^{(5)}=[0,1;0,2)$, $\lambda^{(6)}=[0,2;0,3)$, $\lambda^{(7)}=[0,3;0,5)$, $\lambda^{(8)}=[0,5;0,7)$, $\lambda^{(9)}=[0,7;1)$, $\lambda^{(10)}=[1;\infty)$; для случайных величин $t_{j,j+1}$, $t_{\min}$, $t_{\max}$, характеризующих длины интервалов между эпизодами, $t^{(1)}=[0;0,1)$, $t^{(2)}=[0,1;1)$, $t^{(3)}=[1;7)$, $t^{(4)}=[7;30)$, $t^{(5)}=[30;180)$, $t^{(6)}=[180;\infty)$. Пересчет вероятностей осуществляется с помощью библиотеки Smile [8]. Для автоматизации работы и проведения вычислений реализована программа на языке C# [5], а также ряд скриптов на языке R [7].

Оценивание интенсивности поведения

Описание тестовых данных

Отметим, что, к сожалению, организация исследования, в рамках которого можно собрать как данные об эпизодах поведения, так и фактические сведения об интенсивности поведения, является сложной и требует длительного времени (а в некоторых случаях и невозможна). Поэтому для тестирования модели была разработана программа, генерирующая «эпизоды поведения» в соответствии с теоретическими предположениями модели.

Сначала генерируются 300 значений интенсивности, соответствующие значениям случайной величины, распределенной по гамма-распределению с параметрами $k=1,5$, $\theta=0,15$. Математическое ожидание такой случайной величины 0,225, значения сконцентрированы

в промежутке $(0; 0, 4)$, что соответствует наиболее частым значениям интенсивности в реальном поведении (в частности, если исследуется рискованное поведение в эпидемиологии — частота сексуальных контактов, частота употребления алкоголя и т.д.).

Далее для каждого значения интенсивности генерируется 20 «респондентов» — последовательностей точек, расстояния между которыми подчиняются экспоненциальному распределению с соответствующим значением интенсивности. Из каждой такой последовательности выделяются исходный данные для оценки: длины интервалов между тремя последними точками, минимальный и максимальный интервал за промежуток длиной 183 «дня», удаляются последовательности, у которых нет хотя бы двух точек за этот промежуток. Таким образом, конечный тестовый набор включает 5907 «респондентов», причем для каждого известно исходное значение интенсивности, что позволяет сравнить его с итоговой оценкой.

Интенсивность индивидуального поведения

Одним из применений разработанной модели является оценивание интенсивности поведения индивида. Другими словами, после внесения в модель ответов респондента на вопросы о трех последних эпизодах поведения, а также о минимальном и максимальном интервалах между эпизодами за определенный промежуток времени, происходит пересчет вероятностей. В качестве оценки интенсивности рассматривается интервал с наибольшей вероятностью в апостериорном распределении интенсивности.

Таким образом, задача оценивания интенсивности индивидуального поведения является задачей классификации, в частности, для указанной дискретизации случайной величины λ — классификации по 10 непересекающимся классам. Результаты оценивания интенсивности поведения индивидов по тестовым данным, описанным в предыдущем разделе, представлены в таблице 1. Средняя точность оценивания для 10-классовой классификации 88,9

Интенсивность поведения группы в целом

Отметим, что чаще в исследованиях необходимо вычислять оценку интенсивности для целой группы, или популяции. Например, в случае мониторинга интенсивности незащищенных половых контактов как одного из видов поведения с высоким риском передачи или получения

	Оценка интенсивности									
Исходное значение	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$	$\lambda^{(10)}$
$\lambda^{(1)}$	2	4	0	0	0	0	0	0	0	0
$\lambda^{(2)}$	23	114	48	22	0	0	0	0	0	0
$\lambda^{(3)}$	10	109	103	88	4	0	0	0	0	0
$\lambda^{(4)}$	1	126	263	550	127	6	5	2	0	0
$\lambda^{(5)}$	0	5	35	514	1022	219	140	25	0	0
$\lambda^{(6)}$	0	0	0	20	372	221	311	76	0	0
$\lambda^{(7)}$	0	0	0	1	86	130	457	166	0	0
$\lambda^{(8)}$	0	0	0	0	6	24	182	168	0	0
$\lambda^{(9)}$	0	0	0	0	0	3	26	91	0	0
$\lambda^{(10)}$	0	0	0	0	0	0	0	0	0	0

Таблица 1: Соответствие исходной и оцененной интенсивности

ВИЧ-инфекции, важным является значительное изменение интенсивности изучаемого поведения в целом в группе. Такое изменение может быть, например, показателем успешности проведенных мероприятий (обучения, внедрения социальной рекламы и т.д.), в то время как изменение интенсивности поведения только одного человека может быть вызвано личными причинами и не позволит сделать какие-либо выводы об эффекте, оказанном интервенцией.

Предложенная модель на основе байесовской сети доверия позволяет получать комбинированную оценку за счет свойств распределения Дирихле, являющегося сопряжённым априорным распределением к мультиномиальному распределению [6]. Получение апостериорного распределения интенсивности в группе основывается на сложении векторов вероятностей индивидуальных распределений и последующей их нормировке [3]. Исходное распределение значений интенсивности, согласно которому генерировались данные представлено на рис. 2, финальное, соответствующее оцененным интенсивностям — на рис. 3. Согласно тесту χ^2 мы не отклоняем гипотезу о совпадении этих распределений ($\chi^2 = 80$, $df = 72$, p-value=0.24). Если же говорить о содержательных выводах из апостериорного распределения оценок интенсивности, то можно сказать, что гистограммы имеют сходные формы, большая часть значений сосредоточена в промежутке $(0,05; 0,5)$.

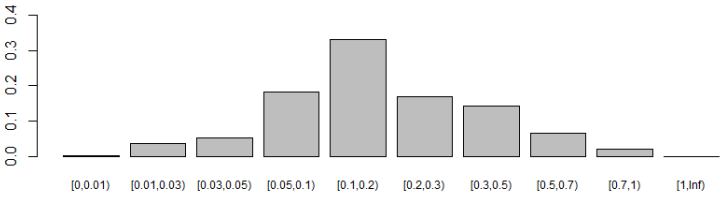

Рис. 2: Исходное распределение интенсивности в группе.

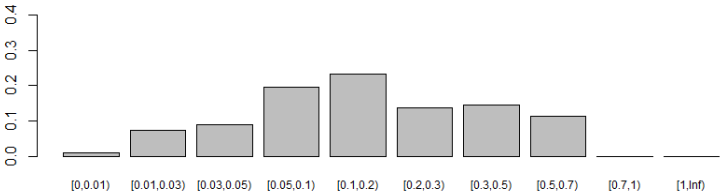

Рис. 3: Распределение оценок интенсивности в группе.

Заключение

Приведенные примеры показывают хорошую согласованность вычисленного распределения интенсивности и первоначального значения, как для индивидуальных, так и для групповых оценок. Однако соответствие теоретической модели реальному поведению требует дополнительного изучения.

Литература

[1] Плавинский С.Л., Баринова А.Н., Разнатовский К.И. Сексуальное поведение, венерические болезни и гетеросексуальная эпидемия ВИЧ-инфекции — некоторые результаты математического моделирования // Российский семейный врач. 2007. Т. 11. N3. С. 30–38.

[2] Суворова А.В. Моделирование социально-значимого поведения по сверхмалой неполной совокупности наблюдений // Информационно-измерительные и управляющие системы. 2013. N9. Т. 11. С. 34–38.

[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] 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.

[5] Хайбуллин Р.Р., Суворова А.В., Тулупьев А.Л. Приложение для синтеза байесовской сети доверия по данным об эпизодах рискованного поведения индивида // Нечеткие системы и мягкие вычисления (НСМВ-2014): труды Шестой всероссийской научно-практической конференции (г. Санкт-Петербург, 27–29 июня 2014 г.). В 2 т. Т. 2. СПб.: Политехника-сервис, 2014. С. 233–239.

[6] Frigyik B.A., Kapila A., Gupta M.R. Introduction to the Dirichlet distribution and related processes. UWEE Tech. Rep. UWEETR-2010-0006. Washington: UWEE, 2010. 27 p.

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

[8] SMILE Engine // BAYESFUSION, LLC. Data Analytics, Mathematical Modeling, Decision Support. URL: http://www.bayesfusion.com

[9] 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.