УДК 004.891 + 311.2
doi:10.15217/issn1684-8853.2018.1.116

СИНТЕЗ СТРУКТУР БАЙЕСОВСКОЙ СЕТИ ДОВЕРИЯ ДЛЯ ОЦЕНКИ ХАРАКТЕРИСТИК РИСКОВАННОГО ПОВЕДЕНИЯ

А. В. Суворова[а], *канд. физ.-мат. наук, старший научный сотрудник, suvalv@gmail.com*
А. Л. Тулупьев[а, б], *доктор физ.-мат. наук, доцент, alexander.tulupyev@gmail.com*
[а]*Санкт-Петербургский институт информатики и автоматизации РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*
[б]*Санкт-Петербургский государственный университет, Университетская наб., 7–9, Санкт-Петербург, 199034, РФ*

Постановка проблемы: *необходимость оценивания параметров поведения (как индивидуального, так и на уровне популяции) возникает в различных областях социологических, психологических, эпидемиологических, маркетинговых исследований, исследований безопасности. Однако прямая оценка интенсивности рискованного поведения не всегда доступна, как следствие, требуется развитие косвенных методов оценки. Ранее был предложен подход к моделированию рискованного поведения на основе байесовской сети доверия по данным о нескольких последних эпизодах такого поведения, но для практического применения необходимы изменения этой модели в целях снижения ее зависимости от первоначальных предположений экспертов о взаимосвязях между элементами модели.* ***Цель:*** *предложить модификацию модели, которая не требует задания структуры экспертами; провести сравнение этой модели с первоначальной.* ***Методы:*** *для проверки модели разработана программа, генерирующая тестовые данные в соответствии с теоретическими предположениями модели. Для построения структуры байесовской сети доверия по сгенерированным данным использован алгоритм оптимизации меры качества сети Hill-Climbing с мерой качества Bayesian Information Criterion.* ***Результаты:*** *предложено развитие подхода к построению модели рискованного поведения на основе байесовской сети доверия по совокупности наблюдений, включающей сведения об эпизодах такого поведения. Проведено сравнение двух структур такой модели: предложенной экспертами и построенной автоматически по данным. В то время как формальные меры качества показывают преимущество автоматически обученной структуры, качество предсказания лучше у модели с экспертно заданной структурой. Таким образом, для решения практических задач можно использовать любую из предложенных моделей; выбор может быть обусловлен условием конкретной задачи.*

Ключевые слова *— моделирование поведения, байесовские сети доверия, синтез структур, машинное обучение, рискованное поведение.*

Цитирование: Суворова А. В., Тулупьев А. Л. Синтез структур байесовской сети доверия для оценки характеристик рискованного поведения // Информационно-управляющие системы. 2018. № 1. С. 116–122. doi:10.15217/issn1684-8853.2018.1.116
Citation: Suvorova A. V., Tulupyev A. L. Bayesian Belief Network Structure Synthesis for Risky Behavior Rate Estimation. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2018, no. 1, pp. 116–122 (In Russian). doi:10.15217/issn1684-8853.2018.1.116

Введение

Решение многих задач в различных областях исследования основано на изучении поведения человека или группы, в частности, на изучении численных параметров исследуемого поведения, сравнении их с нормативными значениями или отслеживании динамики показателей. Примерами подобных задач являются оценивание частоты индивидуального поведения, создающего угрозу безопасности, например информационной системы [1]; моделирование поведения в группе [2]; оценивание интенсивности рискованного поведения на уровне популяции [3].

Ряд задач эпидемиологии и охраны общественного здоровья тесно связан с оценкой вреда, который может быть причинен индивидом обществу, самому себе и (или) другому индивиду. В таком случае с риском связывают эпизоды определенного поведения индивида, а численной характеристикой такого риска выступает интенсивность поведения [4]. Одной из таких задач является оценивание риска передачи или получения ВИЧ-инфекции, а в качестве интересующего исследователя типа поведения — употребление инъекционных наркотиков и рискованное сексуальное поведение [5]. Однако прямая оценка интенсивности (т. е. непосредственное измерение числа эпизодов) рискованного поведения не всегда доступна в силу экономических причин (подобные исследования для сбора данных сложно, долго и дорого организовывать) [6, 7]. Ретроспективное измерение (опросы самих респондентов о числе эпизодов поведения определенного типа) сталкивается с проблемой точности ответов из-за особенностей воспоминания и самооценки, а также с проблемой целенаправленного занижения интенсивности из-за социальной нежелательности многих видов

рискованного поведения [7]. Вот почему требуется развитие косвенных методов оценки.

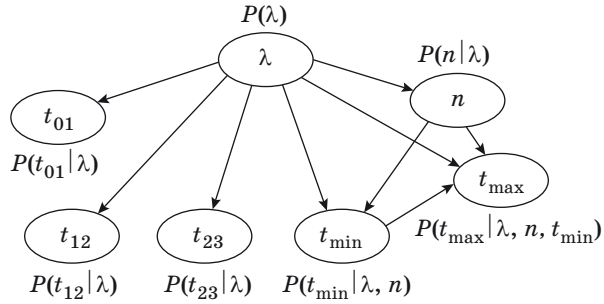
В работах [8, 9] предложен подход к построению байесовской сети доверия (БСД) для моделирования рискованного поведения на основе данных о нескольких последних эпизодах такого поведения. БСД позволяет представить сложные взаимосвязи между элементами, входящими в модель, в виде разложения на более простые элементы, что упрощает как описание всей системы, так и ее интерпретацию [9–11], в то время как выводы делаются с учетом полной системы. Кроме того, аппарат БСД позволяет учитывать как имеющиеся статистические данные, так и экспертную информацию об интересующей исследователя области [9–11]. Причем уровень вовлечения экспертов может быть очень разным: от выбора методов, используемых для построения модели, до полного задания модели, включая ее структуру и параметры [12–15].

В данной работе исследовано влияние экспертных предположений о структуре модели на ее качество. Для этого предложена модификация модели, при построении структуры которой использован алгоритм автоматического обучения по данным, а затем проведено сравнение моделей с исходной и модифицированной структурой.

Описание исходной модели

В качестве исходных данных для вычисления оценки интенсивности рискованного поведения [16] используются сведения о трех последних эпизодах поведения t_{01}, t_{12}, t_{23} (точнее, длины интервалов между этими эпизодами) и сведения о минимальном и максимальном интервале ($t_{\min}$ и $t_{\max}$) между эпизодами за исследуемый промежуток времени. Моделью поведения выступает пуассоновский случайный процесс, т. е. длины интервалов между эпизодами распределены экспоненциально. Кроме того, модель содержит также оцениваемую величину λ, соответствующую интенсивности поведения, и скрытую переменную n — число эпизодов, произошедших за исследуемый промежуток времени.

Для построения байесовской сети возможные значения всех непрерывных величин (λ, t_{01}, t_{12}, t_{23}, $t_{\min}$, $t_{\max}$) разбиваются на дискретные интервалы; таким образом, распределения являются мультиномиальными. 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, $t^{(1)} = [0; 0{,}1)$, $t^{(2)} = [0{,}1; 1)$, $t^{(3)} = [1; 7)$, $t^{(4)} = [7; 30)$, $t^{(5)} = [30; 180)$, $t^{(6)} = [180; \infty)$.



■ ***Рис. 1.*** Структура БСД для моделирования рискованного поведения, заданная экспертно

■ ***Fig. 1.*** Expert-based structure of the BBN for risky behavior modelling

В первоначальной модели [8] структура задана экспертно (рис. 1).

Тензоры **P** условной вероятности, характеризующие переходы между узлами сети: $\mathbf{P} = \left\{P\left(t_{j,j+1}|\lambda\right), P\left(t_{01}|\lambda\right), P\left(t_{\min}|n,\lambda\right), P\left(t_{\max}|n,\lambda,t_{\min}\right), P\left(n|\lambda\right), P\left(\lambda\right)\right\}$, — определяются аналитически, согласно предположению о пуассоновском процессе в качестве модели поведения ($l_s = 1, \ldots, k_s$, где k_s — число дизъюнктных промежутков при дискретизации случайных величин; $s = 0, \ldots, 4$; $j = 1, 2$; $i = 1, \ldots, m$, где m — число дизъюнктных промежутков при дискретизации величины λ) [8]:

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

$$t_{j,j+1}^{(l_j)} = [a; b);$$

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

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

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

$$\times\left(\left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}} - e^{-\lambda^{(i)}b}\right)^{n-1} - \left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}} - e^{-\lambda^{(i)}a}\right)^{n-1}\right),$$

$$t_{\max}^{(l_4)} = [a; b).$$

Предложенная модель была апробирована на тестовых данных [17]; показано, что модель имеет хорошие показатели качества при сравнении с фактическими значениями интенсивности поведения, т. е. сведения о последних эпизодах поведения позволяют характеризовать поведение в целом.

Обучение структуры байесовской сети доверия

Для оценивания влияния экспертных предположений о структуре на качество модели в данной работе также исследуется структура, полученная автоматически по данным. Таблицы условных вероятностей также вычисляются на основе данных. Из-за того, что данные о числе эпизодов реального поведения за заданный промежуток времени получить организационно сложно, а иногда и невозможно (например, при исследовании рискованного сексуального поведения), то для тестирования модели была разработана программа, генерирующая «эпизоды поведения» в соответствии с теоретическими предположениями модели, т. е. в соответствии с пуассоновской моделью поведения. Все вычисления и анализ выполнены с помощью языка R [18], в частности, для работы с байесовскими сетями использовался пакет bnlearn [19].

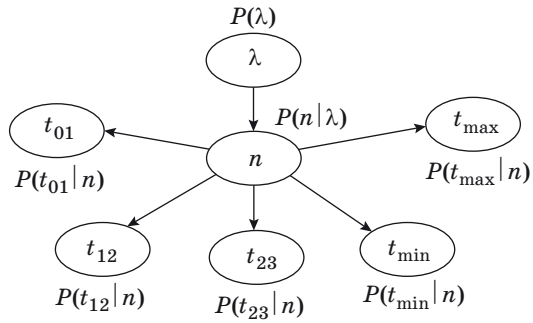
Сначала были сгенерированы 300 значений интенсивности, соответствующие значениям случайной величины, распределенной по гамма-распределению с параметрами $k = 1{,}1$; $\theta = 0{,}3$. С одной стороны, большая часть значений меньше 0,5, что соотносится со многими примерами реального поведения, с другой — в данных есть значения для всех интервалов, на которые разбито значение λ при дискретизации.

Далее для каждого значения интенсивности генерируется 20 «респондентов» — последовательностей 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, конечный обучающий набор включает 5936 «респондентов», причем для каждого дополнительно известно исходное значение интенсивности, что позволяет сравнить его с итоговой оценкой.

Аналогичным образом были сгенерированы тестовые данные для дальнейшей оценки построенных моделей (50 значений интенсивности, 15 последовательностей для каждого значения, всего 730 «респондентов»).

Для построения структуры БСД по сгенерированным данным был использован алгоритм оптимизации меры качества сети Hill-Climbing [19], мерой качества являлся BIC (Bayesian Information Criterion).

Полученная в результате обучения структура БСД представлена на рис. 2.



■ ***Рис. 2.*** Структура БСД для моделирования рискованного поведения, обученная на данных

■ ***Fig. 2.*** Data-based structure of the BBN for risky behavior modelling

Следует отметить, что данная структура имеет достаточно простую интерпретацию: интенсивность поведения (другими словами, частота поведения) определяется через число эпизодов поведения, произошедших за рассматриваемый промежуток времени (в текущем примере 365 дн.). Однако, как уже отмечалось, получить в явном виде число эпизодов для большинства примеров поведения невозможно. Таким образом, n является скрытой переменной, определяемой через исходные данные и определяющей в свою очередь искомое значение λ.

Для полного определения моделей (см. рис. 1, 2) далее было проведено автоматическое обучение параметров представленных байесовских сетей на имеющихся данных, т. е. вычисление условных вероятностей для всех пар переменных, соединенных в разработанных структурах ребром.

Сравним полученные модели. Согласно алгоритму построения, на обучающей выборке мера качества структуры, представленной на рис. 2, выше, чем первоначальной, заданной экспертно (см. рис. 1), как для BIC (–40165 и –54991 соответственно), так и для меры максимального правдоподобия (–38649 и –38704). На тестовой выборке меры качества структуры, обученной по данным, также выше, хотя и незначительно в случае меры максимального правдоподобия (BIC: –5990 vs –16836; максимальное правдоподобие: –4474 vs –4477).

Однако, так как основное назначение предложенных моделей — оценивание интенсивности, то следующий этап сравнения моделей состоит в оценке качества не самих структур, а предсказаний согласно каждой из предложенных моделей. Соответствие предсказанного и исходного значений интенсивности для предложенных моделей представлено в табл. 1 и 2. Отметим, что при указанной дискретизации переменной λ задача оценивания интенсивности индивидуального поведения является задачей классификации по 10 непересекающимся классам.

■ ***Таблица 1.*** Предсказание по экспертно заданной структуре

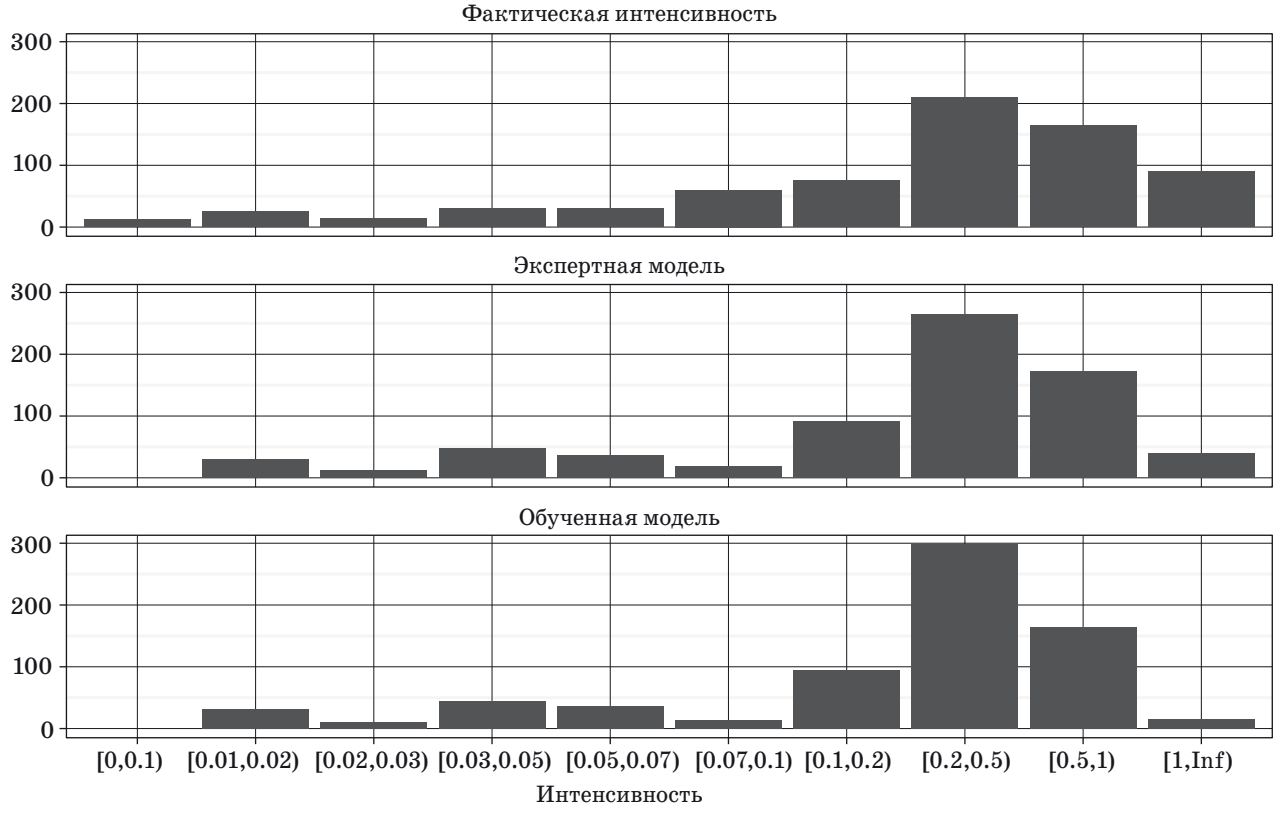
■ ***Table 1.*** Confusion matrix for prediction on model with expert-based structure

		Оценка интенсивности									
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$	$\lambda^{(10)}$
Исходное значение	$\lambda^{(1)}$	0	10	1	1	0	0	0	0	0	0
	$\lambda^{(2)}$	0	14	4	6	1	0	0	0	0	0
	$\lambda^{(3)}$	0	4	3	4	3	0	0	0	0	0
	$\lambda^{(4)}$	0	2	3	17	4	3	1	0	0	0
	$\lambda^{(5)}$	0	1	1	9	11	4	5	0	0	0
	$\lambda^{(6)}$	0	0	0	10	16	11	23	0	0	0
	$\lambda^{(7)}$	0	0	0	1	1	0	48	24	1	0
	$\lambda^{(8)}$	0	0	0	0	0	0	14	162	34	0
	$\lambda^{(9)}$	0	0	0	0	0	0	0	77	81	7
	$\lambda^{(10)}$	0	0	0	0	0	0	0	2	56	32

■ ***Таблица 2.*** Предсказание по обученной на данных структуре

■ ***Table 2.*** Confusion matrix for prediction on model with data-based structure

		Оценка интенсивности									
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$	$\lambda^{(10)}$
Исходное значение	$\lambda^{(1)}$	0	24	2	0	0	0	0	0	0	0
	$\lambda^{(2)}$	0	18	6	3	2	0	0	0	0	0
	$\lambda^{(3)}$	0	4	2	8	1	0	0	0	0	0
	$\lambda^{(4)}$	0	5	0	18	3	2	2	0	0	0
	$\lambda^{(5)}$	0	0	1	9	10	3	7	0	0	0
	$\lambda^{(6)}$	0	0	0	7	20	8	25	0	0	0
	$\lambda^{(7)}$	0	0	0	0	1	0	47	27	0	0
	$\lambda^{(8)}$	0	0	0	0	0	0	14	172	24	0
	$\lambda^{(9)}$	0	0	0	0	0	0	0	94	70	1
	$\lambda^{(10)}$	0	0	0	0	0	0	0	5	71	14



■ ***Рис. 3.*** Распределения интенсивности, полученные на разных моделях

■ ***Fig. 3.*** Predicted rate distributions for different models

Средняя доля правильных предсказаний (accuracy) для 10-классовой классификации согласно модели со структурой, заданной экспертно, немного выше, чем точность оценивания согласно модели с автоматически обученной структурой (90,7% vs 89,8%). Другие показатели качества предсказания также немного выше для экспертной модели (точность (precision): 0,53 vs 0,49; полнота (recall): 0,5 vs 0,48). Отметим, что и в той и в другой модели неправильно классифицируемые значения в большинстве случаев отнесены к интервалам, смежным исходному значению, т. е. даже неправильное предсказание не дает сильного смещения оценки.

Обобщенную оценку по группе в целом можно получить за счет свойств распределения Дирихле,

являющегося сопряженным априорным распределением к мультиномиальному распределению [20]. Апостериорное распределение интенсивности поведения в группе вычисляется путем сложения векторов вероятностей индивидуальных распределений и последующей их нормировки [21]. На рис. 3 представлено фактическое распределение значений интенсивности и распределения, полученные с помощью двух рассматриваемых моделей.

Для рассматриваемых тестовых данных не выявлено различий между распределением интенсивности, посчитанной по первоначальной (экспертной) модели, и распределением интенсивности, вычисленной по модели с автоматически обученной структурой ($\chi^2 = 13{,}86$, $df = 8$, p-value = 0,09). Другими словами, обе модели дают значимо не различающиеся предсказания интенсивности поведения в группе.

Заключение

В работе предложено развитие подхода к построению модели рискованного поведения на основе байесовской сети доверия по совокупности наблюдений, включающей сведения об эпизодах такого поведения. Первоначальная структура модели основана на экспертных знаниях, что может влиять на качество модели. Для снижения подобного влияния рассмотрена другая структура модели, построенная с помощью алгоритма машинного обучения, выявляющего статистические взаимосвязи между элементами из данных и не использующего предположения экспертов о таких взаимосвязях. Результаты сравнения этих структур на автоматически генерируемых данных показывают, что качество предсказания немного выше у модели с экспертно заданной структурой. Формальные меры качества структуры (BIC и мера максимального правдоподобия) ожидаемо выше у структуры, обученной автоматически, так как используемые алгоритмы структурного обучения ориентированы именно на оптимизацию меры качества BIC.

Отметим, однако, что в целом качество предсказания обеих моделей достаточно высокое и не значительно отличается друг от друга. Кроме того, содержательная интерпретация взаимосвязей является обоснованной как для исходной, так и для обученной структуры. Одним из дальнейших направлений исследования является сравнение приведенных в статье моделей не на автоматически генерируемых данных, а на реальных данных о поведении (для такого тестирования, однако, необходимо получить сведения не только о последних эпизодах поведения, но и о фактической интенсивности, что часто невозможно для реального поведения).

Таким образом, на данном этапе для решения практических задач можно использовать любую из предложенных моделей. Выбор может быть обусловлен условием конкретной задачи: например, при наличии некоторой обучающей выборки для оценивания параметров можно использовать структуру, приведенную на рис. 2, а при отсутствии таких данных — первоначальную структуру с аналитически вычисленными условными вероятностями. Перечисленные возможные действия аналитика учитываются при разработке инструментария, автоматизирующего решение задач оценивания параметров поведения.

Статья содержит материалы исследований, частично поддержанных грантом РФФИ 16-31-60063-мол_а_дк.

Литература

1. **Азаров А. А., Тулупьева Т. В., Фильченков А. А., Тулупьев А. Л.** Вероятностно-реляционный подход к представлению модели комплекса «информационная система — персонал — критичные документы» // Тр. СПИИРАН. 2012. Вып. 1(20). С. 57–71. doi:10.15622/sp.20.3
2. **Афанасьев И. В.** Возможности математического моделирования поведения аудитории с помощью динамических математических моделей // Актуальные проблемы современной науки. 2006. № 4. С. 212–218.
3. **Leigh B. C., Stall R.** Substance use and Risky Sexual Behavior for Exposure to HIV: Issues in Methodology, Interpretation, and Prevention // American Psychologist. 1993. Vol. 48(10). P. 1035.
4. **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.
5. **Lemelin C., Lussier Y., Sabourin S., Brassard A., Naud C.** Risky Sexual Behaviours: The Role of Substance use, Psychopathic Traits, and Attachment Insecurity Among Adolescents and Young Adults in Quebec // The Canadian Journal of Human Sexuality. 2014. Vol. 23(3). P. 189–199. doi:0.3138/cjhs.2625
6. **Bolger N., Davis A., Rafaeli E.** Diary Methods: Capturing Life as it is Lived // Annual Review of Psychology. 2003. Vol. 54(1). P. 579–616. doi:10.1146/annurev.psych.54.101601.145030
7. **Graham C. A., Catania J. A., Brand R., Duong T., Canchola J. A.** Recalling Sexual Behavior: A Methodological Analysis of Memory Recall Bias via Interview using the Diary as the Gold Standard // Journal of Sex Research. 2003. Vol. 40(4). P. 325–332. doi:10.1080/00224490209552198

8. **Суворова А. В.** Моделирование социально-значимого поведения по сверхмалой неполной совокупности наблюдений // Информационно-измерительные и управляющие системы. 2013. Т. 11. № 9. С. 34–37.
9. **Тулупьев А. Л., Николенко С. И., Сироткин А. В.** Байесовские сети: логико-вероятностный подход. — СПб.: Наука, 2006. — 607 с.
10. **Neapolitan R. E.** Learning Bayesian Networks. — Pearson Prentice Hall, 2003. — 674 p.
11. **Pearl J.** Causality: Models, Reasoning, and Inference. — Cambridge: Cambridge University Press, 2000. — 400 p.
12. **Constantinou A. C., Fenton N., Marsh W., Radlinski L.** From Complex Questionnaire and Interviewing Data to Intelligent Bayesian Network Models for Medical Decision Support // Artificial Intelligence in Medicine. 2016. P. 75–93. doi:10.1016/j.artmed.2016.01.002
13. **Du Y., Guo Y.** Evidence Reasoning Method for Constructing Conditional Probability Tables in a Bayesian Network of Multimorbidity // Technology and Health Care. 2015. Vol. 23(s1). P. S161–S167. doi:10.3233/thc-150950
14. **Mkrtchyan L., Podofillini L., Dang V. N.** Bayesian Belief Networks for Human Reliability Analysis: A Review of Applications and Gaps // Reliability Engineering & System Safety. 2015. Vol. 139. P. 1–16. doi:10.1016/j.ress.2015.02.006
15. **Trucco P., Cango E., Ruggeri F., Grande O.** A Bayesian Belief Network Modelling of Organisational Factors in Risk Analysis: A Case Study in Maritime Transportation // Engineering and System Safety. 2008. Vol. 93. P. 845–856. doi: 10.1016/j.ress.2007.03.035
16. **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.
17. **Suvorova A., Tulupyeva T.** Bayesian Belief Networks in Risky Behavior Modelling // Advances in Intelligent Systems and Computing. 2016. Vol. 451. P. 95–102. doi:10.1007/978-3-319-33816-3_10
18. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/ (дата обращения: 20.10.2016).
19. **Scutari M.** Learning Bayesian Networks with the Bnlearn R Package // arXiv preprint. arXiv:0908.3817. 2009.
20. **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.
21. **Суворова А. В., Тулупьев А. Л., Сироткин А. В.** Байесовские сети доверия в задачах оценивания интенсивности рискованного поведения // Нечеткие системы и мягкие вычисления. 2014. № 2. С. 115–129.

UDC 004.891 + 311.2
doi:10.15217/issn1684-8853.2018.1.116

Bayesian Belief Network Structure Synthesis for Risky Behavior Rate Estimation

Suvorova A. V.[a], PhD, Phys.-Math., suvalv@gmail.com
Tulupyev A. L.[a,b], Dr. Sc., Phys.-Math., Associate Professor, alexander.tulupyev@gmail.com
[a]Saint-Petersburg Institute for Informatics and Automation of the RAS, 39, 14 Line, V. O., 199178, Saint-Petersburg, Russian Federation
[b]Saint-Petersburg State University, 7–9, Universitetskaya Emb., 199034, Saint-Petersburg, Russian Federation

Introduction: Studies in sociology, psychology, epidemiology, marketing or information security often face the issue of estimating the behavior rate (either individually or at the level of a population). Direct methods of behavior rate estimation are sometimes not available; hence, it is important to develop indirect methods. Earlier studies proposed an approach to risky behavior modeling based on a Bayesian belief network using the data about several last behavior episodes as its initial data. However, to apply this model in practice, we need to reduce its dependency from the initial experts' assumptions about the relations between the elements of the model. **Purpose:** To propose a model modification which would not require an expert-based model structure, and to compare the modified model with the initial one. **Methods:** To test the model, we used an automatically generated dataset which followed some initial assumptions about the data. To form the structure of a Bayesian belief network, we used a score-based hill-climbing algorithm with Bayesian information criterion score. **Results:** We proposed to modify the approach to risky behavior modeling in terms of Bayesian belief network based on the data about several last behavior episodes. The initial expert-based model and the model with a data-based structure were compared. Formal scores were better for the data-based structure, while the prediction quality was slightly better for the expert-based model. Hence, we can use both these models for practical applications; the choice depends on the assumptions and limitations of a particular task.

Keywords — Behavior Modelling, Bayesian Belief Network, Structure Synthesis, Machine Learning, Risky Behavior.

Citation: Suvorova A. V., Tulupyev A. L. Bayesian Belief Network Structure Synthesis for Risky Behavior Rate Estimation. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2018, no. 1, pp. 116–122 (In Russian). doi:10.15217/issn1684-8853.2018.1.116

References

1. Azarov A. A., Tulupyeva T. V., Filchenkov A. A., Tulupyev A. L. Probabilistic Relational Approach to Representing "Informational System — Personnel — Critical Documents" Complex Model. *Trudy SPIIRAN* [SPIIRAS Proceedings], 2012, no. 1(20), pp. 57–71 (In Russian). doi:10.15622/sp.20.3
2. Afanasyev I. V. Capabilities of Dynamic Mathematical Models for Audience Behavior Modeling. *Aktual'nye problemy sovremennoi nauki*, 2006, no. 4, pp. 212–218 (In Russian).
3. Leigh B. C., Stall R. Substance use and Risky Sexual Behavior for Exposure to HIV: Issues in Methodology, Interpretation, and Prevention. *American Psychologist*, 1993, no. 48(10), pp. 1035.
4. 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, no. 29(1), pp. 38–43.
5. Lemelin C., Lussier Y., Sabourin S., Brassard A., Naud C. Risky Sexual Behaviours: The Role of Substance use, Psychopathic Traits, and Attachment Insecurity Among Adolescents and Young Adults in Quebec. *The Canadian Journal of Human Sexuality*, 2014, no. 23(3), pp. 189–199. doi:0.3138/cjhs.2625
6. Bolger N., Davis A., Rafaeli E. Diary Methods: Capturing Life as it is Lived. *Annual Review of Psychology*, 2003, no. 54(1), pp. 579–616. doi:10.1146/annurev.psych.54.101601.145030
7. Graham C. A., Catania J. A., Brand R., Duong T., Canchola J. A. Recalling Sexual Behavior: A Methodological Analysis of Memory Recall Bias via Interview using the Diary as the Gold Standard. *Journal of Sex Research*, 2003, no. 40(4), pp. 325–332. doi:10.1080/00224490209552198
8. Suvorova A. V. Socially Significant Behavior Modeling on the Base of Super-Short Incomplete Set of Observations. *Informatsionno-izmeritel'nye i upravliaiushchie sistemy*, 2013, no. 9(11), pp. 34–37 (In Russian).
9. Tulupyev A. L., Nikolenko S. I., Sirotkin A. V. *Baiesovskie seti: logiko-veroiatnostnyi podkhod* [Bayesian Networks: A Probabilistic Logic Approach]. Saint-Petersburg, Nauka Publ., 2006. 607 p. (In Russian).
10. Neapolitan R. E. *Learning Bayesian Networks*. Pearson Prentice Hall, 2003. 674 p.
11. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge, Cambridge University Press, 2000. 400 p.
12. Constantinou A. C., Fenton N., Marsh W., Radlinski L. From Complex Questionnaire and Interviewing Data to Intelligent Bayesian Network Models for Medical Decision Support. *Artificial Intelligence in Medicine*, 2016, pp. 75–93. doi:10.1016/j.artmed.2016.01.002
13. Du Y., Guo Y. Evidence Reasoning Method for Constructing Conditional Probability Tables in a Bayesian Network of Multimorbidity. *Technology and Health Care*, 2015, no. 23(s1), pp. S161–S167. doi:10.3233/thc-150950
14. Mkrtchyan L., Podofillini L., Dang V. N. Bayesian Belief Networks for Human Reliability Analysis: A Review of Applications and Gaps. *Reliability Engineering & System Safety*, 2015, no. 139, pp. 1–16. doi:10.1016/j.ress.2015.02.006
15. Trucco P., Cango E., Ruggeri F., Grande O. A Bayesian Belief Network Modelling of Organisational Factors in Risk Analysis: A Case Study in Maritime Transportation. *Engineering and System Safety*, 2008, no. 93, pp. 845–856. doi:10.1016/j.ress.2007.03.035
16. Tulupyeva T. V., Pashhenko A. E., Tulupyev A. L., Krasnoselskih T. V., Kazakova O. S. *Modeli VICh-riskovannogo povedeniia v kontekste psikhologicheskoi zashchity i drugikh adaptivnykh stilei* [HIV Risky Behavior Models in the Context of Psychological Defense and Other Adaptive Styles]. Saint-Petersburg, Nauka Publ., 2008. 140 p. (In Russian).
17. Suvorova A., Tulupyeva T. Bayesian Belief Networks in Risky Behavior Modelling. *Advances in Intelligent Systems and Computing*, 2016, no. 451, pp. 95–102. doi:10.1007/978-3-319-33816-3_10
18. *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/ (accessed 20 October 2016).
19. Scutari M. Learning Bayesian Networks with the Bnlearn R Package. *arXiv preprint*. arXiv:0908.3817, 2009.
20. 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.
21. Suvorova A. V., Tulupyev A. L., Sirotkin A. V. Bayesian Belief Networks for Risky Behavior Rate Estimates. *Nechetkie sistemy i miagkie vychisleniia* [Fuzzy Systems and Soft Computing], 2014, no. 2, pp. 115–129 (In Russian).