Подходы к обработке зашумленных данных в модели социально-значимого поведения

А. В. Торопова[1], А. В. Суворова[2]
Санкт-Петербургский институт информатики и автоматизации РАН
[1]alexandra.toropova@gmail.com, [2]suvalv@gmail.com

***Аннотация*. Модель социально-значимого поведения основана на данных, получаемых от респондентов. Такие данные могут быть неточными, что негативно сказывается на результатах работы модели. Для работы с такими данными в докладе предлагаются следующие подходы: основанный на согласованности данных и на использовании в модели скрытых переменных, отвечающих за реальные данные.**

Ключевые слова: байесовские сети доверия; БСД; байесовская сеть доверия со скрытыми переменными; модель социально-значимого поведения; зашумленные данные

I. Введение

Во многих задачах социологии, психологии и других наук, изучающих человека и его поведение, возникает необходимость оценки интенсивности социально-значимого поведения человека (т.е. такого поведения, которое оказывает какое-либо влияние на других людей). Для измерения этого параметра в работах [1–3] была предложена модель на основе байесовской сети доверия, благодаря тому, что байесовская сеть доверия позволяет представлять данные с неполнотой удобным для вычислений способом, а также позволяет определять апостериорные распределения, входящих в модель случайных элементов при появлении новых данных о значении наблюдаемых величин [4].

Работа модели построена на данных респондентов о последних трех, а также минимальном и максимальном интервалах между эпизодами исследуемого поведения. В связи с тем, что исследуемое поведение может быть социально-неодобряемым, респонденты могут дать априори ложные данные, чтобы интенсивность их поведения казалась меньшей (или же наоборот большей в случае социально-одобряемого поведения). Кроме того, из-за работы человеческой памяти респонденты могут несознательно ошибиться в своих ответах. Таким образом, оценка интенсивности, вычисленная с помощью модели социально-значимого поведения может оказаться неточной.

Для увеличения точности модели мы предлагаем следующие два подхода к обработке искаженных данных: первый основан на согласованности исходных данных, а второй использует скрытые переменные, характеризующие реальные значения интервалов между эпизодами поведения респондентов, а не их ответы.

Работа выполнена в рамках проекта по государственному заданию СПИИРАН № 0073-2018-0001

II. Модель социально-значимого поведения

На рис. 1 (модель создана в редакторе GeNIe [5]) модель $M=\left(G(V,L),\mathbf{P}\right)$ представлена байесовской сетью доверия [4]. Структура модели представлена графом $G(V,L)$, где $V=\left\{t_{01},t_{12},t_{23},t_{\min},t_{\max},\lambda,n\right\}$ – множество вершин, $L=\left\{(u,v):u,v\in V\right\}$ — множество направленных связей между вершинами.

Rate – это случайная величина, характеризующая интенсивность поведения, t_{ij} – случайная величина, характеризующая длину интервала между *i*-ым и *j*-ым с конца эпизодами. Также в модель включены минимальный и максимальный интервалы между эпизодами ($t_{\min}$ и $t_{\max}$ соответственно). n – случайная величина, характеризующая число эпизодов за исследуемый период.

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

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

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

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

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

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

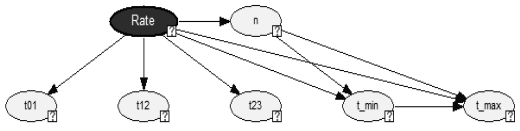

Рис. 1. Модель социально-значимого поведения

III. Зашумленные данные

В данном докладе мы используем зашумленные данные сгенерированные автоматически.

Для начала генерируются данные об интервалах между эпизодами в соответствии с теоретическими предположениями модели социально-значимого поведения.

После этого на промежутках, зависящих от каждого из интервалов, мы выберем случайные величины таким образом, что отклонение между моментом интервью и последним эпизодом будет не более четверти, между последним и предпоследним – не более половины, а между предпоследним и предпредпоследним будет отличаться менее, чем в два раза.

К минимальному и максимальному интервалам мы добавим нормальный шум.

Такой подход к генерации зашумленных данных имеет смысл, поскольку находит отражение в жизни: чем раньше произошло какое-то событие, в данном случае эпизод социально-значимого поведения, тем сложнее вспомнить, когда именно оно произошло.

Для дальнейшей работы было сгенерирована 2 сета: 746 ответов респондентов для тестирования подходов и 5998 для обучения модели со скрытыми переменными.

IV. Подходы к работе с зашумленными данными

Мы предлагаем два возможных варианта работы с зашумленными данными. Расскажем о каждом из них подробнее.

A. Согласованность данных

На рис. 2 представлена модель социально-значимого поведения, дополненная узлами, позволяющими произвести диагностику согласованности данных полученных от респондента.

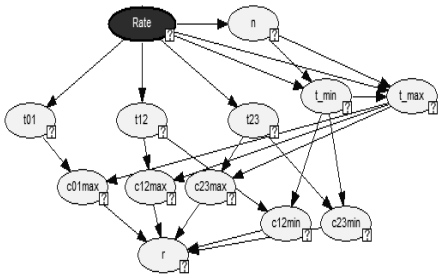

Рис. 2. Модель социально-значимого поведения, расширенная аппаратом диагностики согласованности данных

Вершины $c_{t_{12,\min}}$ и $c_{t_{23,\min}}$ показывают степень согласованности эпизода t_{ij} с минимальным интервалом $t_{\min}$, вершины $c_{t_{01,\max}}$, $c_{t_{12,\max}}$ и $c_{t_{23,\max}}$ эпизода t_{ij} – с максимальным интервалом $t_{\max}$. $c_{t_{01,\min}}$ не рассматривается, так как t_{01} представляет собой интервал между моментом интервьюирования, который не является эпизодом исследуемого поведения, и последним эпизодом поведения. Оценка согласованности $c_{t_{ij,\min}}$ может принимать значения: $c^{+}_{t_{ij,\min}}$ (t_{ij} и $t_{\min}$ согласованы), $c^{-}_{t_{ij,\min}}$ (t_{ij} и $t_{\min}$ не согласованы) и $c^{?}_{t_{ij,\min}}$ (t_{ij} и $t_{\min}$ находятся в одном и том же интервале). Тензоры условной вероятности, характеризующие переходы к добавленному узлу в общем случае определяются следующим образом:

$$p\left(c^{(s)}_{t_{ij,\min}} \mid t_{ij}, t_{\min}\right) = \begin{cases} \alpha^{(s)}, & t_{ij} > t_{\min}; \\ \beta^{(s)}, & t_{ij} < t_{\min}; \\ 1-\alpha^{(s)}-\beta^{(s)}, & t_{ij} = t_{\min}; \end{cases}$$

где $s \in \{+,-,?\}$, $\alpha^{(s)}, \beta^{(s)} \in [0;1]$, $\alpha^{(s)} + \beta^{(s)} \le 1$, $\sum \alpha = 1$, $\sum \beta = 1$. Аналогичным образом рассматривается оценки согласованности $c_{t_{ij,\max}}$.

Вершина r характеризует оценку надежности респондента в целом. Чтобы упростить формулы для условных вероятностей.

Обозначим $c = \left(c_{t_{12,\min}}, c_{t_{23,\min}}, c_{t_{01,\max}}, c_{t_{12,\max}}, c_{t_{23,\max}}\right)$, тогда $p(r^{+} \mid c) = \dfrac{\sum c^{+}}{\sum c}$, $p(r^{-} \mid c) = \dfrac{\sum c^{-}}{\sum c}$ и $p(r^{?} \mid c) = \dfrac{\sum c^{?}}{\sum c}$.

После того как зашумленные данные были обработаны с помощью такой диагностики, задача исследователя определить с какими данными продолжать работу. Допустим, что для исследования нужно использовать данные только тех респондентов, у которых нет противоречий или каких-либо неопределенностей в их ответах (данные респондента полностью согласованы), тогда из сета данных 746 респондентов останутся данные только 278-ми респондентов.

B. Скрытые переменные

К описанной модели социально-значимого поведения были добавлены вершины t^0_{01}, t^0_{12}, t^0_{23}, $t^0_{\min}$ и $t^0_{\max}$ (рис. 3), представляющие интервалы поведения, полученные из ответов респондентов (то же самое, что и t_{01}, t_{12}, t_{23}, $t_{\min}$ и $t_{\max}$ в исходной модели). А t_{01}, t_{12}, t_{23}, $t_{\min}$ и $t_{\max}$ теперь это скрытые переменные, которые характеризуют действительные последние интервалы поведения. Дело в том, что в ответах респондентов может содержаться неточная или даже заведомо неправильная информация (такое может произойти, например, из-за желания одобрения поведения респондента), то есть реальные интервалы неизвестны, даны только ответы респондентов.

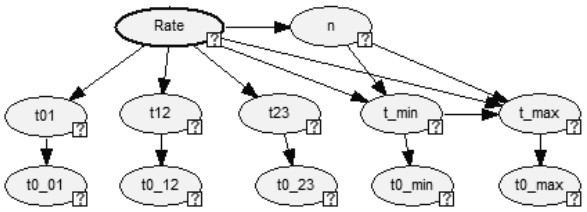

Рис. 3. Модель социально-значимого поведения со скрытыми переменными

С помощью сгенерированного для обучения сета с зашумленными данными эта модель была обучена.

Сравним теперь модель, расширенную скрытыми переменными и исходную модель социально-значимого поведения, используя сгенерированные для тестирования моделей данные ответов респондентов.

Таблицы соответствия предсказанного и исходного значений интенсивности для предложенных моделей представлены в табл. 1 (модель, расширенная скрытыми переменными) и в табл. 2 (исходная модель социально-значимого поведения). Отметим, что при указанной дискретизации переменной λ, задача оценивания интенсивности индивидуального поведения является задачей классификации по 10 непересекающимся классам.

ТАБЛИЦА I ПРЕДСКАЗАНИЕ МОДЕЛИ СО СКРЫТЫМИ ПЕРЕМЕННЫМИ

		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	3	1	1	0	1	0	0	0	0
	$\lambda^{(2)}$	0	8	7	5	3	3	1	0	0	0
	$\lambda^{(3)}$	0	0	2	4	2	2	1	0	0	0
	$\lambda^{(4)}$	0	0	1	7	7	12	3	0	0	0
	$\lambda^{(5)}$	0	0	1	7	10	26	16	0	0	0
	$\lambda^{(6)}$	0	0	0	0	0	0	0	0	0	0
	$\lambda^{(7)}$	0	0	1	3	5	26	50	63	2	0
	$\lambda^{(8)}$	0	0	0	0	0	2	18	223	12	0
	$\lambda^{(9)}$	0	0	0	0	0	0	1	113	47	4
	$\lambda^{(10)}$	0	0	0	0	0	0	0	1	20	9

ТАБЛИЦА II ПРЕДСКАЗАНИЕ ИСХОДНОЙ МОДЕЛИ

		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	9	2	0	0	0	0	0	0	0
	$\lambda^{(2)}$	0	8	6	9	5	0	1	0	0	0
	$\lambda^{(3)}$	0	3	3	7	0	0	1	0	0	0
	$\lambda^{(4)}$	0	0	0	13	3	13	1	0	0	0
	$\lambda^{(5)}$	0	0	2	9	5	25	18	1	0	0
	$\lambda^{(6)}$	0	0	0	0	0	0	0	0	0	0
	$\lambda^{(7)}$	0	0	0	4	1	31	77	37	0	0
	$\lambda^{(8)}$	0	0	0	0	0	14	78	156	7	0
	$\lambda^{(9)}$	0	0	0	0	0	3	25	89	33	1
	$\lambda^{(10)}$	0	0	0	0	0	0	0	0	7	9

Средняя точность оценивания (accuracy) для 10-классовой классификации согласно модели со скрытыми переменными немного выше (89,7%), чем точность оценивания исходной модели (88,5%).

V. ЗАКЛЮЧЕНИЕ

Для работы с зашумленными данными было предложено два подхода, каждый из них может быть использован в зависимости от вида проводимого исследования.

Было показано, что модель со скрытыми переменными дает более точные предсказания по сравнению с исходной моделью социально-значимого поведения.

Полученные результаты могут быть использованы в различных научных областях, предметом исследования которых является поведение человека, например, в социологии и эпидемиологии.

СПИСОК ЛИТЕРАТУРЫ

[1] Суворова А.В. Модели и алгоритмы анализа сверхкоротких гранулярных временных рядов на основе байе-совских сетей доверия. Диссертация на соискание ученой степени кандидата физико-математических наук. 2013

[2] Суворова А.В. Моделирование социально-значимого поведения по сверхмалой неполной совокупности наблюдений // Информационно-измерительные и управляющие системы. 2013. №9, 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, № 2. С. 115-129

[4] Тулупьев А.Л., Сироткин А.В., Николенко С.И. Байесовские сети доверия: логико-вероятностный вывод в ациклических направленных графах. СПб.: Изд-во С.-Петерб. ун-та, 2009. 400 с.

[5] GeNIe&SMILE // URL: http://download.bayesfusion.com/files.html?category=Academia (дата обращения 01.04.2018)