Оценка среднего числа эпизодов постинга в онлайн медиа по цифровым следам пользователя: неполным данным о времени публикаций и характеристикам профиля

В. Ф. Столярова
Санкт-Петербургский Федеральный исследовательский центр РАН
vfs@dscs.pro

А. Л. Тулупьев[1,2]
[1]*Санкт-Петербургский государственный университет;*
[2]*Санкт-Петербургский Федеральный исследовательский центр РАН*
alt@dscs.pro

***Аннотация*. В ряде областей знаний, которые тесно связаны с поведением человека, возникает задача оценки связанного с ним риска, причем доступные данные чаще всего неполны и неточны. В качестве характеристики, отражающей паттерн эпизодического поведения индивида, в работе рассмотрено среднее число эпизодов за промежуток времени. Для гамма-пуассоновской модели поведения представлены подходы для моделирования кумулятивной средней функции числа эпизодов, в том числе с учетом наблюдаемых и ненаблюдаемых характеристик индивида. Работа предложенного подхода к оценке риска представлена на данных о публикации постов в онлайн медиа.**

Ключевые слова: социоинженерные атаки; кумулятивная средняя функция числа эпизодов; последние эпизоды; гамма-пуассоновская модель поведения

I. Введение

Деятельность человека в современных социокиберфизических системах рассматривается как источник риска и включается в систему анализа рисков организации [10]. Согласно [13], значительная доля несчастных случаев в химической промышленности, морских и авиаперевозках происходит от ошибок персонала. В социоориентированных областях знаний, таких как здравоохранение, также отмечается рост, связанных с ошибками сотрудников [4].

Человеческий фактор играет одну из главных ролей также в реализации угроз информационной безопасности предприятия [6, 16]. При этом отдельной задачей является оценка защищенности пользователей информационной системы от социоинженерных атакующих действий, направленных на несанкционированный доступ к критичной информации [16, 25].

Информация о поведении пользователя информационной системы складывается из нескольких источников: самоотчеты о поведении, получаемые в рамках специализированных интервью и опросов; экспертные мнения; а также цифровой портрет пользователя. Отмечается, что цифровые следы пользователя, в том числе частота постинга в онлайн медиа [5], отражают психологические особенности его личности [8, 16], и потому тесно связаны с риском успешной реализации социоинженерной атаки [18].

Кроме того, существует непосредственно рискованное поведение, связанное с кибербезопасностью, как, например, передача паролей третьим лицам или переход по ссылке из письма с незнакомого адреса. Для получения данных о подобном рискованном поведении используются интервью и самоотчеты пользователей информационной системы. Например, в работах [7, 9] были предложены различные шкалы для оценки осведомленности об информационной безопасности. В основе опросных инструментариев лежат утверждения о различных типах рискованного (с точки зрения кибербезопасности) поведения, для которых пользователю предлагается оценить степень согласия (от полностью согласен до полностью не согласен). Работа [1] посвящена безопасности использования онлайн медиа, исследование опирается, в том числе, на вопросы о частоте смены пароля профиля и частоте обновления информации. В работе [14] используются также вопросы о частоте обновления настроек безопасности профиля онлайн медиа.

Однако такая частотная информация, полученная в результате интервью, часто подвержена влиянию когнитивных искажений. Обращение к наиболее запомнившимся эпизодам рискованного поведения помогает снизить эту неопределенность. В работе [20] был предложен метод оценки частоты поведения по датам нескольких последних последовательных эпизодов для поведения, связанного с риском передачи ВИЧ. Вопросы о последних эпизодах поведения с одной стороны позволяют обратиться к наиболее запомнившейся информации, и с другой затруднительны для ответов в контексте социальной ожидаемости [24]. Однако такая неполная информация не может напрямую использоваться для определения числовых характеристик по таким неполным необходимо привлечение математических моделей. Классической базой для них служат точечные случайные процессы, широко используемые в приложениях [2]. В простейшем случае используется

Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН № FFZF-2022-0003 и при финансовой поддержке РФФИ грант №20-07-00839

процесс Пуассона [21, 19], если же такая модель не может описать имеющуюся в выборке дисперсию, то интенсивность процесса может полагаться случайной, варьирующейся между индивидами. В этом случае говорят о гамма-пуассоновской модель поведения [20, 17].

Математический аппарат для обработки неполной информации о нескольких последовательных эпизодах поведения развивался в ряде работ [20, 24, 21, 17, 23, 11]. В работе [22] была предложена формализация гамма-пуассоновской модели поведения, позволяющая использовать методы анализа времени жизни при оценке характеристик поведения, в частности, регрессию Кокса для определения функции интенсивности случайного процесса. Функция интенсивности случайного процесса полностью определяет его характеристики, однако в том случае, если данных немного, требуются более робастные методы.

Целью работы является формализация задачи оценки кумулятивной средней функции числа эпизодов некоторого рискованного поведения, а также регрессионной модели для определения степени влияния внешних факторов на постинг в онлайн медиа. Использование средней функции вместо функции интенсивности в качестве зависимой переменной позволяет отказаться от ряда предположений модели и получать робастные оценки [2, 3]. Предлагаемые модели в области анализа данных об эпизодах поведения из интервью и опросов, являются новыми.

II. Средняя функция числа эпизодов: формализация задачи

Основной моделью процесса реализации эпизодов поведения на временной оси выступает точечный случайный процесс. Рассмотрим выборку n индивидов, каждый индивид i предоставляет информацию о n_i эпизодах некоторого поведения за промежуток времени $[\tau_{i1}, \tau_{i2}]$, например, о постах в онлайн медиа или о разглашении пароля третьим лицам, которые происходят в моменты времени $t_{i1},\ldots,t_{in_i}$. Обозначим $N_i(t)$ – число эпизодов, реализованных индивидом i к моменту времени t. Точечный случайный процесс характеризуется средним значением числа эпизодов $\mu_i(t) = E(N_i(t))$ и их дисперсией σ_i.

Для поведения человека естественно выполнены следующие предпосылки: все эпизоды происходят в непрерывном времени, за конечное время может произойти конечное число эпизодов, два эпизода не могут произойти одновременно. Таким образом, базовой моделью для реализации эпизодов отдельного индивида может считаться пуассоновский процесс (возможно, неоднородный).

Если почти для всех индивидов известны данные лишь об одном–двух эпизодах поведения, то полная вероятностная спецификация модели посредством функции интенсивности случайного процесса может оказаться недостаточно точной [2, 3]. В этом случае для получения более робастных оценок оправдан переход к заданию кумулятивной средней функции случайного процесса. Такая функция отражает зависимость среднего числа эпизодов поведения, реализованных в популяции, от времени. Вид такой функции (а именно выпукла она или вогнута) указывает на тенденции в поведении для рассматриваемой выборки.

A. *Выборочная оценка кумулятивной средней функции Нельсона–Аалена*

Кумулятивная средняя функция может быть построена непараметрическим способом. Если процессы повторяющихся событий для всех индивидов $N_i(t)$ в выборке независимы и имеют одинаковую среднюю функцию $\mu(t)$, то кумулятивная средняя функция $M(t) = \int_0^t m(s)ds$ может быть оценена (оценка максимума правдоподобия) как

$$\hat{M}(t) = \int_0^t \frac{\sum_{j=1}^{n} dN_j(s)}{\sum_{j=1}^{n} \delta_j},$$

где в каждый момент s наблюдается δ_j индивидов, и произошло $dN_j(s)$ событий. Такая оценка кумулятивной средней функции является несмещенной и состоятельной оценкой истинного значения $M(t)$ в том случае, если моменты остановки наблюдений не зависят от самих процессов реализации эпизодов [3].

B. *Регрессионная модель для учета внешних факторов*

Чтобы учесть влияние на среднюю функцию числа эпизодов различных индивидуальных факторов, как наблюдаемых (пол, возраст и т. д.), так и ненаблюдаемых, как индивидуальная склонность к поведению, обуславливающая различия в интенсивности поведения от пользователя к пользователю в гамма-пуассоновской модели поведения [22], используется регрессионная модель. Пусть на поведение каждого индивида i оказывают влияние вектор внешних наблюдаемых факторов $x_i(t)$ и некоторая ненаблюдаемая склонность к поведению u_i, которая моделируется гамма-распределенной случайной величиной. Пусть также для всех индивидов в выборке все u_i независимы и одинаково распределены со средним 1. Рассмотрим функцию интенсивности случайного процесса в виде [2]

$$\lambda_i(t \mid u) = u_i \lambda_0(t) \exp(\beta^T x_i(t)), \qquad (1)$$

где β есть вектор коэффициентов регрессии, а $\lambda_0(t)$ — базовая функция интенсивности. Такая функция полностью определяет процесс реализации эпизодов на временной оси. Учитывая, что u_i имеет среднее значение 1, средняя функция имеет вид:

$$\mu(t) = \int_0^t \exp(\beta^T x_i(u)) d\mu_0(u).$$

Таким образом, используя подгонку регрессии (1), можно также строить оценки кумулятивной средней функции.

III. Оценка среднего числа эпизодов постинга в онлайн медиа

Для демонстрации работы предложенного подхода обратимся к данным о постинге в онлайн социальной сети. Частота публикации постов является одним из цифровых следов пользователя информационной системы.

Были собраны данные о постах 1000 пользователей онлайн медиа «ВКонтакте» за один календарный год с момента последнего посещения интернет-ресурса, имеющих активные профили. В качестве момента окончания исследования (момента цензурирования) использовалась дата последнего посещения своей страницы. Таким образом, в качестве τ_{i2} выбирался момент последнего посещения ресурса, а в качестве момента $\tau_{i1} = \tau_{i2} - 365.25$. Дальнейшая подгонка кумулятивной средней функции осуществлялась с помощью пакета reda статистической среды обработки данных R [15].

На рис. 1 представлена выборочная оценка Нельсона–Аалена кумулятивной средней функции эпизодов постинга в выборке. Выпуклый характер функции свидетельствует о том, что кумулятивное среднее выше ближе к концу периода, то есть в выборке участвовало некоторое число наблюдений, опубликовавших пост незадолго до момента последнего посещения.

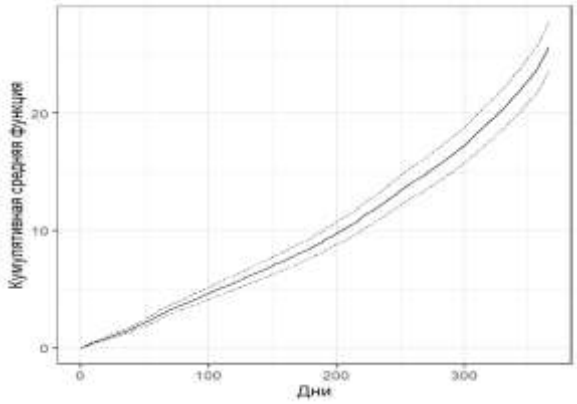

Рис. 1. Выборочная кумулятивная функция эпизодов постинга в выборке

При подгонке регрессионной модели (1), были получены значения коэффициентов регрессии, представленные в таблице.

Фактор	Коэфф. регрессии	Стандартное отклонение	p-значение
Возраст	0.017	0.004	<0.01
Пол (мужской)	-0.439	0.128	<0.01
Число друзей (логарифм+1)	0.151	0.037	<0.01

Отметим, что переменная «количество друзей» была преобразована. Оценка параметра ненаблюдаемой переменной 0.674 (стандартное отклонение 0.042).

Чтобы проиллюстрировать влияние рассматриваемых факторов на кумулятивную среднюю функцию, была сделана выборка из ещё 1000 пользователей онлайн медиа. Для нового тестового набора данных было сравнено кумулятивное число эпизодов постинга по полу, указанному в профиле онлайн медиа. Согласно отрицательному коэффициенту при соответствующей переменной в регрессионной модели, если указан мужской пол, то кумулятивное число эпизодов меньше. На рис. 2 представлены кумулятивные средние значения эпизодов для двух значений пола.

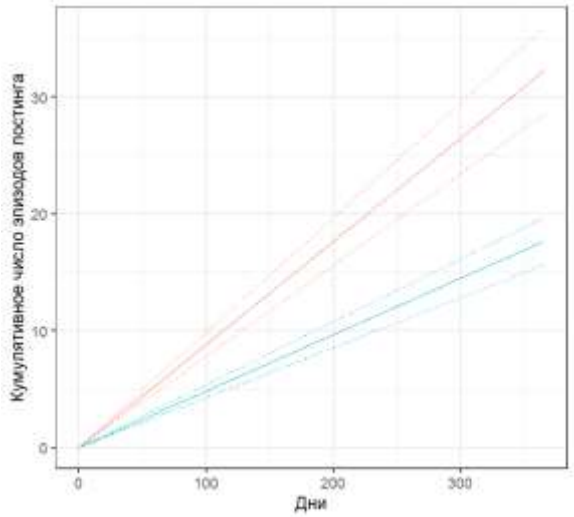

Рис. 2. Сравнение кумулятивных средних функций числа эпизодов для мужского и женского пола

IV. Заключение

В ряде случаев оправданным как с точки зрения доступного набора данных, так и с точки зрения интерпретации результатов является обращение к кумулятивной средней функции, отражающей общее количество эпизодов, которые произошли в выборке за определенный период времени. С математической точки зрения для моделирования такой ситуации используется методы анализа времени жизни: оценка Нельсона–Аалена и регрессия типа Кокса [2].

В работе представлена формализация задачи оценки кумулятивного числа эпизодов в рамках анализа эпизодического, связанного с риском поведения индивида. Отметим, что кумулятивное число эпизодов поведения в популяции в ряде моделей связано напрямую с риском, как, например, в области эпидемиологии при распространении ВИЧ-инфекции [24]. В других областях эти величины связаны опосредованно, например, частота передачи пароля третьи лицам влечет повышенный риск успешной реализации социоинженерной атаки [16].

Для данных о постинге в онлайн медиа представлены как выборочная оценка кумулятивной средней функции числа эпизодов за один год, так и регрессионная модель, учитывающая наблюдаемые (пол, возраст, число друзей) и ненаблюдаемые (индивидуальная склонность к поведению) характеристики поведения. Предложенные методы являются новыми в области анализа данных о поведении человека.

Список литературы

[1] Acquisti A., Gross R. Imagined communities: Awareness, information sharing, and privacy on the Facebook //International workshop on privacy enhancing technologies. – Springer, Berlin, Heidelberg, 2006. С. 36-58.

[2] Cook R.J., Lawless J. The statistical analysis of recurrent events // Springer Science and Business Media, 2007. 402 p.

[3] Lawless J. F., Nadeau C. Some simple robust methods for the analysis of recurrent events //Technometrics. 1995. Т. 37, №. 2. С. 158–168.

[4] Evans M.G., He Y., Yevseyeva I., Janicke H. Published incidents and their proportions of human error //Information and Computer Security. 2019. Т. 27, No. 3. С. 343–357.

[5] Junco R. Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance // Computers in human behavior. 2012. Vol. 28. No. 1. Pp. 187–198.

[6] Metalidou E., Marinagi, C., Trivellas, P., Eberhagen, N., Skourlas, C., Giannakopoulos, G. The human factor of information security: Unintentional damage perspective //Procedia-Social and Behavioral Sciences. 2014. Т 147. С. 424–428.

[7] Öğütçü G., Testik Ö. M., Chouseinoglou O. Analysis of personal information security behavior and awareness //Computers & Security. 2016. Т. 56. С. 83-93.

[8] Orosz G., Tóth-Király I., Bőthe B. Four facets of Facebook intensity—the development of the multidimensional Facebook intensity scale //Personality and individual differences. 2016. Т. 100. С. 95-104.

[9] Parsons K., Calic D., Pattinson M., Butavicius M., McCormac A., Zwaans T.The human aspects of information security questionnaire (HAIS-Q): two further validation studies //Computers & Security. 2017. Т. 66. С. 40-51.

[10] Prochazkova D., Prochazka J. Risk-Based Design of Socio-Cyber-Physical Systems //International Journal of Computer and Information Technology (2279–0764). 2021. Т. 10. №. 2.

[11] Stoliarova V.F. Non-parametric Bayes belief network for intensity estimation with data on several last episodes of person's behavior //International Scientific and Practical Conference in Control Engineering and Decision Making. Springer, Cham, 2020. С. 486-497.

[12] Swain A.D. Human reliability analysis: Need, status, trends and limitations // Reliability Engineering & System Safety. 1990. Т. 29. №. 3. С. 301–313.

[13] Tao, J., Qiu, D., Yang, F., Duan, Z. A bibliometric analysis of human reliability research //Journal of Cleaner Production. 2020. Т. 260. С. 121041.

[14] Utz S., Krämer N. The privacy paradox on social network sites revisited: The role of individual characteristics and group norms //Cyberpsychology: Journal of psychosocial research on cyberspace. 2009. Т. 3. №. 2. С. 2.

[15] Wang W, Fu H, Yan J._reda: Recurrent Event Data Analysis_. R package version 0.5.3, <URL: https://github.com/wenjie2wang/reda>, 2021.

[16] Абрамов М.В., Тулупьев А.Л., Тулупьева Т.В. Социоинженерные атаки: социальные сети и оценки защищенности пользователей. 2018. 268 с.

[17] Зельтерман Д., Суворова А.В., Пащенко А.Е., Мусина В.Ф., Тулупьев А.Л., Тулупьева Т.В., Красносельских Т.В., Гро Л.Е., Хаймер Р. Обработка систематической ошибки, связанной с длиной временных интервалов между интервью и последним эпизодом в гамма-пуассоновской модели поведения // Труды СПИИРАН, 2011. Вып. 16. С. 160–185.

[18] Корепанова А.А., Олисеенко В.Д., Абрамов М.В., Тулупьев А.Л. Применение методов машинного обучения в задаче идентификации аккаунтов пользователя в двух социальных сетях //Компьютерные инструменты в образовании. 2019. №. 3. С. 29-43.

[19] Пащенко А.Е., Тулупьев А.Л., Тулупьева Т.В., Красносельских Т.В., Соколовский Е.В. Косвенная оценка вероятности заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения // Здравоохранение Российской Федерации. 2010. № 2. 32–35.

[20] Пащенко А.Е., Тулупьев А.Л., Николенко С.И. Моделирование заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения //Известия высших учебных заведений. Приборостроение. 2006. Т. 49. №. 11. С. 33-34.

[21] Степанов Д.В., Мусина В.Ф., Суворова А.В., Тулупьев А.Л., Сироткин А.В., Тулупьева Т.В. Функция правдоподобия с гетерогенными аргументами в идентификации пуассоновской модели рискованного поведения в случае информационного дефицита // Труды СПИИРАН. 2012. 4(23). С. 157–184.

[22] Столярова В.Ф., Тулупьев А.Л. Регрессия Кокса в задаче оценки параметров рискообразующего поведения индивида по данным о последних эпизодах //Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Физико-математические науки. 2021. Т. 14, №. 4. С. 202-217.

[23] 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.

[24] 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: монография. СПб.: 2008. 346 с.

[25] Хлобыстова А.О., Абрамов М.В., Тулупьев А.Л. Идентификация наиболее вероятных траекторий социоинженерных атак в управлении рисками, ассоциированными с пользователями //Информационные технологии в управлении (ИТУ-2018). 2018. С. 493-496