Регрессия в задаче оценки параметров гамма-пуассоновской модели поведения: апробация на данных о постинге в онлайн социальной сети

В. Ф. Столярова
СПб ФИЦ РАН
vfs@dscs.pro

А. Л. Тулупьев
СПбГУ, СПб ФИЦ РАН
alt@dscs.pro

***Аннотация*. Оценка параметров поведения человека требуется при анализе риска, который связан с деятельностью человека в социальных, технических и информационных системах. Гамма-пуассоновская модель поведения используется в прикладных задачах, где доступна ограниченная информация о поведении человека. В статье представлена регрессионная модель для оценки параметров интенсивности поведения в гамма–пуассоновской модели. Предложенная модель апробирована на данных о публикации постов в онлайн социальной сети ВКонтакте.**

Ключевые слова: гамма-пуассоновская модель поведения, функция правдоподобия, регрессия Кокса, онлайн социальная сеть

I. Введение

Деятельность человека в социальных, технических, информационных, киберфизических, социо-киберфизических системах может быть связана с возможностью реализации некоторого негативного исхода. Негативный исход рассматривается в рамках исследования с точки зрения экономического риска: ущерба благосостоянию индивида, компании или общества. Поэтому оценка параметров поведения человека требуется при анализе риска в подобных системах.

Информация о поведении индивида может быть получена из разных источников. К примеру, онлайн социальные сети играют большую роль в жизни человека, и не только отражают его психологические характеристики [1, 12, 23, 24], но и могут оказывать влияние на поведение человека [6, 7]. Порой их использование может вызывать зависимость [3, 12] или быть индикатором психологического неблагополучия индивида [4]. В этом случае моделирование паттерна поведения человека в онлайн социальной сети важно не только для оценки различных аспектов человеко–компьютерного взаимодействия (таких как частота уведомлений в соответствующих приложениях) [3], но и для оценки риска его здоровью. Паттерны поведения представляют собой модель данных о поведении, и включают в себя ряд характеристик, как внешних (окружение, среда), так и внутренних (индивидуальная склонность к поведению) [22]. Одной из внутренних характеристик эпизодического поведения является его интенсивность, т. е. число эпизодов, которые произошли за определенный промежуток времени.

Паттерны поведения человека в сети интернет могут использоваться работодателями для оценки качеств сотрудника, к примеру, частота использования онлайн социальных сетей связана с академической успеваемостью студентов [11]. Оценка различных характеристик использования онлайн социальных сетей может быть интересна и в области маркетинговых исследований, так как связана с восприятием пользователем брендов и рекламы [14]. Кроме того, поведение сотрудника организации может нести угрозу безопасности критичной информации компании. В частности, сотрудник может является целью *социоинженерной атаки злоумышленника* [17, 22].

Также исследование паттернов поведения человека в онлайн социальной сети может быть важно с точки зрения охраны общественного здоровья. В исследовании [10] отражена взаимосвязь между интенсивностью использования онлайн социальных сетей связана с интенсивностью употребления алкоголя и наркотиков. Отметим, что ответы на вопросы об эпизодах рискованного поведения (внутривенного употребления наркотиков, незащищенных половых актах и т. п.) часто подвержены различным типам когнитивных искажений, например искажение, связанное с ошибками памяти, и социально обусловленные искажения (попытка попасть в социальные ожидания). Чтобы снизить влияние искажений на оценки параметров поведения по данным из интервью и самоотчетов был предложен метод их оценки по данным о последних эпизодах поведения [19, 21].

В связи с возрастающей ролью персонализированных информационных технологий задача моделирования паттернов поведения пользователя онлайн социальной сети является актуальной. Исследователю может быть доступна различная информация о поведении индивида, как из самоотчетов [8], так и открытая информация из профиля пользователя. Обращение к интервью и самоотчетам является ценным источником информации в

Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН № 0073-2019-0003 и при финансовой поддержке РФФИ грант №20-07-00839

социологических и психологических исследованиях, а также если техническая информация недоступна в силу ограничений приватности. В этом интервью и самоотчеты являются единственным источником информации о поведении пользователя в онлайн социальной сети. Кроме того, при сборе информации о паттернах поведения в онлайн социальной сети вручную, для ускорения и удешевления процесса может быть доступна информация лишь о нескольких последних действиях пользователя.

Для построения оценок интенсивности поведения по данным о последних его эпизодах используется байесовские сети доверия [26]. Однако при использовании этого метода возникает задача дискретизации входящих в модель переменных, решение которой в свою очередь зависит от области знаний, в которой требуется оценка параметров поведения и от конкретных данных, используемых в модели [27]. Требуется разработка новых более универсальных подходов к оценке параметров поведения.

В рамках данного исследования предложена регрессионная модель для оценки параметров интенсивности постинга в онлайн социальной сети по данным о нескольких последовательных эпизодах. Регрессионная модель позволяет включать различные детерминанты поведения респондента (внешние характеристики паттерна поведения). Модель апробирована на данных о постинге в онлайн социальной сети ВКонтакте.

II. Функция правдоподобия для оценки параметров гамма–пуассоновской модели поведения

Рассмотрим эпизодическое поведение m индивидов. На протяжении времени каждый индивид совершает некоторые разовые действия, такие как публикация постов. Пусть в результате исследования каждый индивид i наблюдается отрезок времени $(0,\tau_i)$, и за это время произошло n_i последовательных эпизодов поведения. Такая информация может быть получена в результате интервью (самооценка интенсивности поведения) или же в результате обращения к общедоступной информации профиля. Таким образом, эпизодическое поведение индивида представляет собой точечный случайный процесс $N(t)$. Обозначим функцию интенсивности такого процесса:

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{Pr\{\Delta N(t) = 1|H(t)\}}{\Delta t},$$

где $H(t)$ представляет собой историю эпизодов процесса до времени *t*. Функция интенсивности процесса полностью определяет его характеристики [9].

Определим конкретный вид функции интенсивности процесса поведения. Пусть эпизоды поведения каждого индивида происходят согласно вехам процесса Пуассона [5]. В этом случае функция интенсивности случайного процесса имеет вид:

$$\lambda(t|H(t)) = \rho(t), t > 0.$$

В этом случае говорят о *пуассоновской модели поведения* [20].

Однако при моделировании поведения в приложениях возникает необходимость учета различных внешних факторов, таких как пол, возраст, социо-экономический статус респондента. Кроме того, бывает важно учесть ненаблюдаемые индивидуальные характеристики, которые отражают индивидуальную склонность к поведению (individual accident proneness) [9]. Математической моделью таких ненаблюдаемых факторов выступает набор случайных величин u_i. Распространенным и удобным с вычислительной точки зрения является выбор гамма-распределения для u_i [5, 9]. Предположим также, что $u_i, i = 1 \ldots m$ являются независимыми и одинаково распределенными случайными величинами. В этом случае говорят о *гамма-пуассоновской модели поведения*, которая впервые была упомянута в работе [18, 25]. В основе гамма-пуассоновской модели поведения лежит смешанный процесс Пуассона [9].

Зависимость функции интенсивности процесса от внешних факторов (ковариант) $\mathbf{x}_i$ и от *случайного фактора* u_i может иметь различные формы [5, 13]. Интенсивность смешанного процесса Пуассона при наблюдении случайных факторов имеет вид:

$$\lambda_i(t|u_i) = u_i \rho_i(t) = u_i \rho_0(t;\alpha) \exp(\mathbf{x}_i \boldsymbol{\beta}), \quad (1)$$

где $\rho_0(t,\alpha)$ является базовой функцией интенсивности (если все внешние факторы $\mathbf{x}_i$ принимают значение 0), $\mathbf{x}_i$ – вектор внешних ковариант, $\boldsymbol{\beta}$ – вектор коэффициентов регрессии, u_i – случайные величины с гамма-распределением вероятности. Обратим внимание, что в дальнейших рассуждениях без ограничения общности можно считать, что гамма-распределение вероятности имеет среднее 1. Нормировочная константа может быть включена в свободный член регрессии внешних факторов. Таким образом, искомое гамма-распределение вероятности имеет единственный параметр формы распределения ϕ (в случае единичного математического ожидания, дисперсия гамма-распределения равна ϕ):

$$g(u;\phi) = \frac{u^{\phi^{-1}-1}\exp(-u/\phi)}{\phi^{\phi^{-1}}\Gamma(1/\phi)}, u > 0,$$

Такой вид интенсивности точечного случайного процесса является частным случаем *модели пропорциональных рисков Кокса* [16], в которую включена так называемая frailty («хрупкость», склонность индивида к риску). Регрессионный анализ подобных моделей основан на методе максимального правдоподобия [5]. Обозначим вектор параметров модели θ. Он включает в себя параметр базовой функции интенсивности α (при параметрическом задании), вектор коэффициентов регрессии β, параметр гамма-распределения интенсивности поведения ϕ. Общая функция

правдоподобия представляет собой произведение функций правдоподобия реализаций конкретных эпизодов для каждого из m индивидов в выборке:

$$L(\theta) = \prod_{i=1}^{m} L_i(\theta).$$

Если бы u_i наблюдаются, то правдоподобие данных $\{n_i, t_{i1}, \ldots, t_{in_i}, u_i\}$ для каждого индивида *i* выражается как

$$L_i(\theta) = \prod_{j=1}^{n_i} \left(u_i \rho_i(t_{ij}) \right) exp\left\{ \int_0^{\infty} Y_i(s)\ u_i \rho_i(s)\, ds \right\}$$

Здесь в модель включен процесс наблюдения $Y_i(t) = I(t < \tau_i)$, который отражает, наблюдается ли индивид в момент времени *t*, т. е. произошло ли событие в конце интервала наблюдения или нет. Общая функция правдоподобия имеет вид

$$L_C(\theta, \phi) = \prod_{i=1}^{n} \left[\prod_{j=1}^{n_i} \frac{\rho_i(t_{ij})}{\mu_i(\tau_i)} \right] (u_i \mu_i(\tau_i))^{n_i} \exp(-u_i \mu_i(\tau_i)) \times$$

$$\times \frac{u_i^{\phi^{-1}-1} \exp(-u_i/\phi)}{\Gamma(\phi^{-1})\, \phi^{\phi^{-1}}},$$

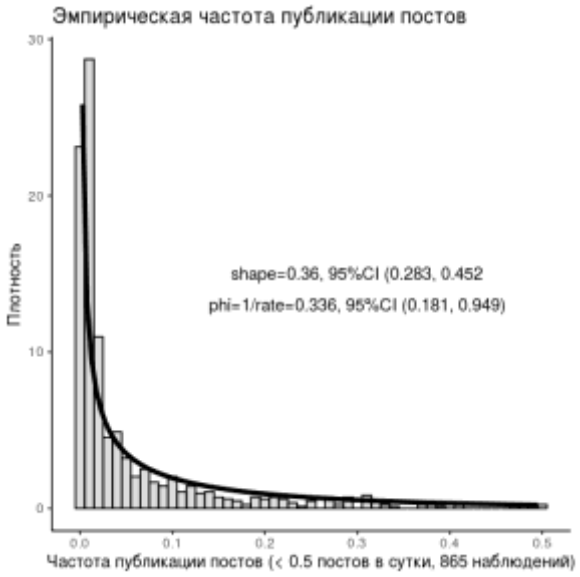
где $\mu_i(t) = \int_0^t \rho_i(s)\, ds$.

В рассматриваемой базовая функция интенсивности $\rho_0(t, \alpha)$ не наблюдается, таким образом, используется полупараметрический вывод оценок максимального правдоподобия. Для построения оценок параметров используется EM-метод.

III. Реализация в среде R

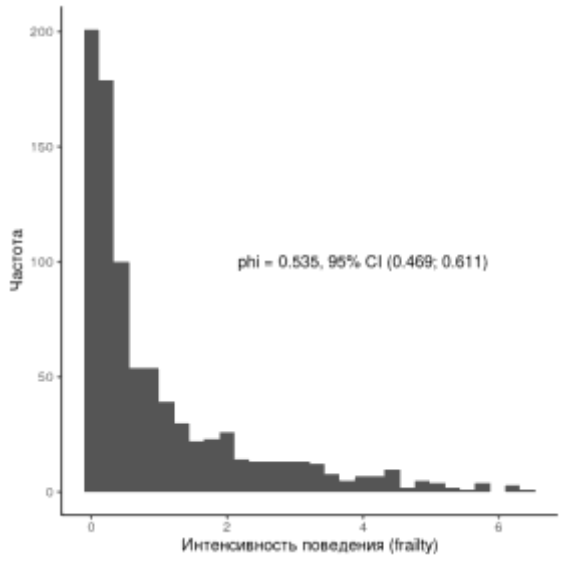
При помощи API Вконтакте были собраны данные о 865 пользователях, которые публиковали посты за последний год (с момента сбора информации). Для каждого пользователя вычислялось количество постов за один календарный год с момента сбора информации. На графике представлена гистограмма частот и соответствующая подгонка гамма-распределения вероятности (в гамма-пуассоновской модели интенсивность поведения в популяции имеет гамма–распределение вероятности). При помощи метода бутсрэп (4000 репликаций исходной выборки) были получены оценки для параметра дисперсии гамма-распределения вероятности:

$\widehat{EmpInt} = 0.336 \in (0.181, 0.949)$.



Для каждого индивида была создана таблица выживаемости: каждая строка содержит данные о начале и конце интервала времени, а так же произошел ли эпизод в конце интервала или нет. Действительно, мы наблюдаем не только полные интервалы между событиями процесса, но и особый интервал между последним эпизодом и моментом сбора информации. Далее была подогнана регрессия Кокса с гамма-распределенным параметром ненаблюдаемой гетерогенности (frailty). Для подгонки использовался пакет emfrail среды обработки данных R [2]. 95 % доверительный интервал для оценки параметра ϕ составляет

$\$ \phi = 0.535 \in (0.469; 0.611) \$$.



IV. Заключение

В ряде ситуаций исследователю доступен лишь ограниченный набор сведений о поведении человека: последние эпизоды. В статье рассмотрен метод оценки параметров интенсивности постинга в онлайн социальной сети по данным о последних эпизодах в рамках гамма–пуассоновской модели поведения. В рамках этой модели предполагается, что интенсивность поведения в популяции представляется гамма-распределенной случайной величиной. Для построения оценки максимального правдоподобия параметра этого гамма-распределения использовалась регрессия Кокса с гамма-распределенным случайным членом (frailty). Полученные при помощи метода доверительные интервалы уже, чем доверительный интервал для эмпирической интенсивности поведения (полученный методом бутстрэп), что является свидетельством применимости метода последних эпизодов для решения поставленной задачи.

Полученный результат является новым в области моделирования паттернов поведения индивида в онлайн социальной сети по данным о последних его эпизодах. Использование регрессии для построения оценок параметров поведения позволяет не только получать оценки параметров поведения, минуя этап дискретизации, но и использовать при оценке различные внешние факторы, оказывающие влияние на поведение. Однако этот подход имеет ограничения к применимости при малом числе наблюдений (индивидов).

Список литературы

[1] Bachrach Y., Kosinski M., Graepel T., Kohli P., Stillwell D. Personality and patterns of Facebook usage // In Proceedings of the 4th annual ACM web science conference. 2012. Pp. 24–32.

[2] Balan T.A. and Putter H. frailtyEM: An R Package for Estimating Semiparametric Shared Frailty Models // Journal of Statistical Software. 2019. No 90(7). Pp. 1–29.

[3] Bedjaoui M., Elouali N., Benslimane S. M. User time spent between persuasiveness and usability of social networking mobile applications: a case study of Facebook and YouTube // Proceedings of the 16th International Conference on Advances in Mobile Computing and Multimedia. 2018. Pp. 15-24.

[4] Charoensukmongkol P., Moqbel M., Gutierrez-Wirsching S. Social media sites use intensity and job burnout among the US and Thai employees // International Journal of Cyber Behavior, Psychology and Learning (IJCBPL). 2017. Vol. 7, No 1. Pp. 34–51.

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

[6] Dhir A., Tsai C.C. Understanding the relationship between intensity and gratifications of Facebook use among adolescents and young adults // Telematics and Informatics. 2017. Vol. 34, No 4. Pp. 350–364.

[7] Ellison N.B., Steinfield C, Lampe C. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites // Journal of Computer-Mediated Communication, Vol.12, № 4, 2007. Pp. 1143–1168.

[8] Gosling S. D., Augustine A. A., Vazire S., Holtzman N., Gaddis, S. Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information // Cyberpsychology, Behavior, and Social Networking. 2011. No 14(9). Pp. 483–488.

[9] Grandell J. Mixed poisson processes. CRC Press, 1997. Monographs in Statistics and Probability 77. 268 p.

[10] Ilakkuvan V., Johnson A., Villanti A.C., Evans W.D., Turner M. Patterns of social media use and their relationship to health risks among young adults // Journal of Adolescent Health. 2019. No 64(2). 158–164.

[11] 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.

[12] Kuss D.J., Griffiths M.D. Online Social Networking and Addiction — A Review of the Psychological Literature // International Journal of Environmental Research and Public Health. 2011. No 8(9). Pp. 3528–3552.

[13] Lawless J. F. Regression methods for Poisson process data // Journal of the American Statistical Association. 1987. Vol. 82. No 399. Pp. 808–815.

[14] Phua J., Ahn S.J. Explicating the 'like' on Facebook brand pages: The effect of intensity of Facebook use, number of overall 'likes', and number of friends' 'likes' on consumers' brand outcomes // Journal of Marketing Communications. 2016. Vol. 22, No 5. Pp. 544–559.

[15] Su C. C., Chan N. K. Predicting social capital on Facebook: The implications of use intensity, perceived content desirability, and Facebook-enabled communication practices // Computers in Human Behavior. 2017. Vol. 72. Pp. 259–268.

[16] Therneau T.M. and Grambsch P.M. Modeling Survival Data: Extending the Cox Model. NewYork:Springer, 2000, ISBN 0-387-98784-3, 350 pp.

[17] Азаров А.А., Тулупьева Т.В., Суворова А.В., Тулупьев А.Л., Абрамов М.В., Юсупов Р.М. Социоинженерные атаки: проблемы анализа. СПб., Наука, 2016. 349 с.

[18] 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.

[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] 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.

[21] 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. (монография) 346 с.

[22] Абрамов М.В., Тулупьев А.Л., Тулупьева Т.В. Социоинженерные атаки: социальные сети и оценки защищенности пользователей. СПб:ГУАП, 2018. 266 с.

[23] Abramov M.V., Azarov A.A. Identifying user's of social networks psychological features on the basis of their musical preferences // Soft Computing and Measurements (SCM), 2017 XX IEEE International Conference on. IEEE, 2017. P. 90–92.

[24] Bagretsov G.I., Shindarev N.A., Abramov M.V., Tulupyeva T.V. Approaches to development of models for text analysis of information in social network profiles in order to evaluate user's vulnerabilities profile // Soft Computing and Measurements (SCM), 2017 XX IEEE International Conference on. IEEE, 2017. P. 93–95.

[25] Stoliarova V. Non-Parametric Bayes Belief Network for Intensity Estimation with Data on Several Last Episodes of Person's Behavior // in Dolinina O. et al. (eds) Recent Research in Control Engineering and Decision Making. ICIT 2020. Studies in Systems, Decision and Control, vol 337. Springer, Cham., p. 486-497.

[26] Suvorova A., Tulupyeva T. Bayesian belief networks in risky behavior modelling // Proceedings of the First International Scientific Conference "Intelligent Information Technologies for Industry"(IITI'16). Springer, Cham, 2016. Pp. 95-102.

[27] Maslove D. M., Podchiyska T., Lowe H. J. Discretization of continuous features in clinical datasets //Journal of the American Medical Informatics Association. 2013. Vol 20. No 3. Pp. 544-553.