Источники неопределенности и способы их обработки в задаче оценки сводных числовых характеристик поведения по данным из самоотчетов индивидов о его последних эпизодах [1]

Столярова В.Ф., м.н.с. лаборатории теоретических и междисциплинарных проблем информатики СПб ФИЦ РАН, svf@dscs.pro

Аннотация

Работа посвящена задаче учета неопределенности ситуации сбора самоотчетов при оценивании кумулятивных характеристик эпизодического поведения индивидов. В работе представлена классификация типов неопределенности, которые возникают в ситуации сбора и анализа данных самоотчетов. Классическая модель эпизодического поведения представлет собой случайный точечный процесс. В работе описаны как классический подход к учету неопределенности в оценке кумулятивных характеристик поведения посредством регресионного анализа, так и байесовский подход на основе байесовских сети доверия, которые позволят строить гибкие модели с учетом особенностей каждой ситуации сбора самоотчетов.

Введение

Сбор и анализ данных о поведении человека является важной частью решения задач в социоориентированных областях знаний. Такие данные используются для описания популяции или выборки, с которой работает специалист, но и для построения прогностических моделей и систем поддержки принятия решений. Подобные системы часто направлены на оценку риска, который связан с исследуемым поведением, и на последующий анализ экономической составляющей различных вариантов развития событий. Например, при решении задачи оценки уровня защищенности пользователей от социоинженерных атакующих

[1] Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН СПИИРАН № FFZF-2022-0003.

воздействий может быть важно выявить степень приверженности индивида рискованным действиям, связанным с информационной системой, как предоставление пароля третьим лицам[1, 2, 3]. Другим примером является исследование обыденного поведения индивидов, которые может быть связано с риском развития хронических заболеваний, как физическая активность [4].

Для извлечения информации о подобном поведении используются как объективные методы, включающие различные мониторы, анализы или прямое наблюдение [4], или же косвенные методы, которые опираются на различные опросы и интервью. Если первый класс методов является ресурсозатратным и часто нереализуемым при исследовании девиантного поведения, то одномоментные самоотчеты о поведении часто недооценивают истинные показатели, в отличие от дневниковых методов, анкет из нескольких пунктов или сбора данных в режиме реального времени[6]. Золотым стандартом при анкетировании или интервьюировании респондентов об их поведении является календарные методы [7, 8]. С развитием индивидуальных информационных технологий, а также доступности интернета, появился еще один способ сбора информации о повседневном поведении человека: EMA (ecological momentary assessment) [5], который подразумевает внесение индивидом данных об эпизодах своего поведения в момент их реализации посредством технических средств. Такой подход комбинирует в себе черты классического опроса и прямого наблюдения, при этом позволяя собирать дополнительную информацию о внешних факторах, сопутствующих поведению.

Однако поведение индивида является многогранным объектом, включающим множество различных факторов и характеристик. При этом при сборе самоотчетов исследователь сталкивается с рядом неопределенностей, связанных с когнитивными особенностями респондента. Поэтому при построении оценок кумулятивных характеристик эпизодического рискованного поведения возникает задача выявления, а также моделирования возникающей неопределенности. Целью исследования является классификация источников неопределенности в ситуации сбора самоотчетов индивидов об эпизодическом поведении, а также способы обработки неопределенности при построении оценок кумулятивных характеристик такого поведения. Новизна исследования в подходе (на основе вероятностных графических моделей) к учету неопределенности. Теоретическая значимость состоит в системном подходе к анализу рисков в ситуации сбора самоотчетов с целью построения оценок характеристик поведения, соответствующие вероятностные графи-

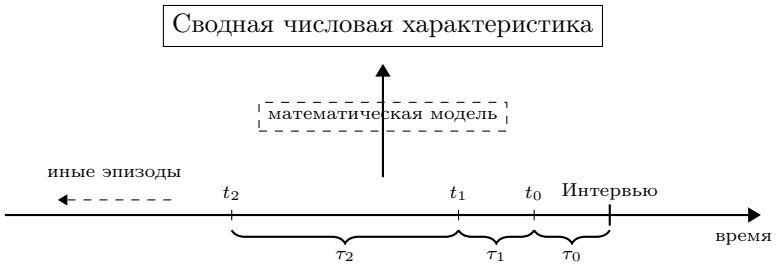

Рис. 1: Математическая постановка задачи оценки кумулятивной характеристики поведения по данным о нескольких последних эпизодах

ческие модели составляют практическую значимость.

Математическая модель эпизодического поведения индивидов

В работах [9, 10, **?**] была предложена и разработана математическая модель эпизодического поведения индивида, которая легла в основу метода построения оценки кумулятивных характеристик такого поведения по неполным и неточным данным об эпизодах. Для построения оценки в модели используются данные о нескольких последовательных эпизодах поведения. Математическим ядром для этой модели служит теория анализа повторяющихся событий [12]. На рисунке 1 представлена схематическое изображение предложенного подхода.

Предположим, что исследуется какое-то эпизодическое поведение. Пусть для каждого индивида i из выборки n извлекается информация о нескольких последовательных эпизодах поведения с помощью некоторого метода $0 < t^i_{m_i} < \ldots < t^i_2 < t^i_1 < t^i_0 < I$ (I - это момент опроса или сбора данных). Это может быть как метод EMA или же стандартное интервью. Обозначим как $N^i(t)$ количество эпизодов в интервале $(0, t]$, а как $H(t) = \{N(s), 0 \leq s \leq t\}$ — историю точечного процесса до времени $t > 0$. Такой точечный процесс может задаваться посредством функции интенсивности, которая отражает мгновенную вероятность реализации эпизода.

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

Такая формализация процесса эпизодического поведения позволяет использовать регерссионный анализ для оценки интенсивности процесса,

которая в свою очередь служит основой для вычисления различных кумулятивных характеристик процесса реализации эпизодов [12].

Однако для практических приложений была предложена модель оценивания искомых характеристик процесса на основе байесовской сети доверия[13, 14]. Такой подход является достаточно гибким, позволяющим моделировать неопределенность из различных источников.

Классификация видов неопределенности и подходы к их обработке

В целом ситуация сбора самоотчетов позволяет выявить следующие источники неопределенности:

1. Неопределенность внешних факторов, возникающая в силу возможной неизвестности факторов, оказывающих влияние на эпизоды поведения, которые могут быть взаимозависимыми.

2. Неопределенность, связанная с неоднородностью поведения в популяции: различные респонденты имеют различную склонность к поведению, а также могут предоставлять информацию о различном числе эпизодов.

3. Неопределенность информации о дате эпизода поведения, которая может поступать искаженной в силу особенностей мышления респондента.

Далее приведен сравнительный анализ двух подходов к построению оценки кумулятивных характеристик поведения с точки зрения учета выяввленных неопределенностей.

Классический подход к анализу повторяющихся событий позволяет учесть неопределенность, связанную с неоднородностью выборки, путем включения случайного множителя u_i, отражающего индивидуальную склонность к поведению, в функции интенсивности процесса. В этом случае функция интенсивности будет иметь вид:

$$\lambda(t|u_i) = u_i \rho_i(t),$$

где $\rho_i(t)$ — функция, отражающая зависимость от времени и фиксированных ковариат, влияющих на процесс генерации эпизода.

Наиболее распространенной и простой с вычислительной точки зрения является выбор гамма–распределения для u_i. Кроме того, число

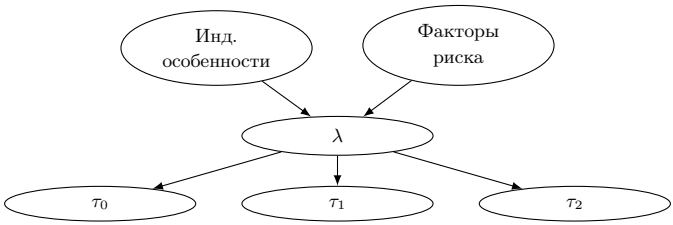

Рис. 2: Расширенная для учета возникающих типов неопределенности структура байесовской сети доверия

эпизодов, наблюдаемое для каждого индивида, может быть различным при квантификации функции интенсивность по статистичсеким данным посредством регрессии (регрессии Кокса [12]). Такая модель широко используется в различных приложениях [12], включая моделирование поведения человека [9, 15]. Внешние факторы, влияющие на поведение, могут быть включены в модель только в форме регрессии [11], что ограничивает возмодности моделирования первого типа неопределенности, связанной со структурой внешних факторов. При этом отметим, что учет неточности в самих самоотчетах, т.е. неточности данных о времени реализации эпизода, с помощью регрессионной модели практически невозможен.

Для учета перечисленных типов неопределенности байесовская сеть доверия, предложенная в работе [13], может быть дополнена соответствующими узлами (см. рисунок 2). В ситуации малого числа наблюдений, байесовские сети доверия позволяют использовать экспертную информацию как для задания структуры сети, что может быть полезно при обработке неопределенности внешних факторов и их структуры зависимости, так и для квантификации модели. Обратим внимание также, что существуют подходы к учету в ьайесовских сетях доверия неопределенности самих данных (тип 3) путем добавления скрытых переменных

Существуют подходы к численному заданию байесовских сетей доверия, которые позволяют использовать непрерывные случаный элементы для байесовского вывода напрямую, минуя этап дискретизации [16].

Заключение

В работе представлена классификация типов неопределенности, которые возникают в ситуации сбора и анализа данных самоотчетов. Классическая модель эпизодического поведения представлет собой случайный точечный процесс. В работе описан классический подход к учету неопределенности в оценке кумулятивных характеристик поведения посредством регресионного анализа. Однако область применимости этого метода ограничена в силу жесткости задания регрессионной модели. Более гибким подходом оказываются байесовские сети доверия. В работе предложена структура байесовской сети доверия для оценивания кумулятивных характеристик поведения в рамках гамма–пуассоновской модели эпизодического поведения индивидов с учетом возникающих источников неопределенности.

Список литературы

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

[2] Khlobystova A., Abramov M. The models separation of access rights of users to critical documents of information system as factor of reduce impact of successful social engineering attacks //CEUR Workshop Proceedings. 2020. T 2782. C. 264-268.

[3] Abramov M. V., Tulupyev A. L. Soft Estimates of User Protection from Social Engineering Attacks // Conference on Artificial Intelligence and Natural Language. – Springer, Cham, 2019. – C. 47-58.

[4] Prince S.A., Adamo K.B., Hamel M.E., Hardt J., Gorber S.C., Tremblay MA comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review //International journal of behavioral nutrition and physical activity. – 2008. – Т. 5. №. 1. С. 1-24.

[5] Shiffman S., Stone A. A., Hufford M. R. Ecological momentary assessment //Annu. Rev. Clin. Psychol. – 2008. – Т. 4. – С. 1-32.

[6] Prince S. A., Cardilli L., Reed J. L., Saunders T. J., Kite C., Douillette K., Fournier K., Buckley J. P. A comparison of self-reported and

device measured sedentary behaviour in adults: a systematic review and meta-analysis //International Journal of Behavioral Nutrition and Physical Activity. – 2020. – Т. 17. – №. 1. – С. 1-17.

[7] Sobell L. C., Sobell M. B. Timeline follow-back //Measuring alcohol consumption. – Humana Press, Totowa, NJ, 1992. – С. 41-72.

[8] Liu W., Li R., Zimmerman M. A., Walton M. A., Cunningham R. M., Buu A. Liu W. et al. Statistical methods for evaluating the correlation between timeline follow-back data and daily process data with applications to research on alcohol and marijuana use // Addictive behaviors. – 2019. – Т. 94. – С. 147-155.

[9] 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.

[10] Tulupyev A., Suvorova A., Sousa J., Zelterman D. Beta prime regression with application to risky behavior frequency screening //Statistics in medicine. – 2013. – Т. 32. – №. 23. – С. 4044-4056.

[11] Stoliarova V., Tulupyev A. L. Cox regression in the problem of risky behavior parameter estimation based on the last episodes' data //St. Petersburg State Polytechnical University Journal. Physics and Mathematics. – 2021. – Т. 14. – №. 4. – С. 202.

[12] Cook R. J. et al. The statistical analysis of recurrent events. – New York : Springer, 2007. – С. 128-133.

[13] 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.

[14] Тулупьев А. Л., Николенко С. И., Сироткин А. В. Основы теории байесовских сетей доверия. – 2019. 399 стр.

[15] Lin T. H., Tsai M. H. Solving unobserved heterogeneity with latent class inflated Poisson regression model //Journal of Applied Statistics. – 2022. – Т. 49. – №. 11. – С. 2953-2963.

[16] Stoliarova V., Tulupyev A. Probabilistic Graphical Models with Continuous Variables for the Decision Making About Risky Episodic

Behavior in the Framework of Gamma Poisson Model with Application to Public Posting Data //International Conference on Intelligent Information Technologies for Industry. – Springer, Cham, 2023. – С. 465-474.