Подходы к учету неопределенности данных о времени реализации эпизодов в моделях оценивания сводных характеристик эпизодического поведения индивидов

В. Ф. Столярова[1], Т. В. Тулупьева[2,1]

[1]*Санкт-Петербургский Федеральный исследовательский центр Российской академии наук*
[2]*Северо-Западный институт управления Российской академии народного хозяйства и государственной службы при Президенте РФ*

vfs@dscs.pro, tvt@dscs.pro

***Аннотация*. Гамма-пуассоновская модель поведения возникает в контексте анализа паттернов эпизодического поведения индивидов для построения оценок сводных характеристик по данным о времени реализации эпизодов поведения. Подобные задачи возникают, в частности, при анализе поведения, ассоциированного с риском. Однако если данные собираются в рамках самоотчета, то они могут являться неточными в силу искажения припоминания или же выраженными на естественном языке. В работе представлены способы учета возникающей неопределенности в различных моделях оценивания искомых характеристик: регрессии и гибридной байесовской сети доверия. Проведены вычислительные эксперименты, проанализирована устойчивость существующих моделей.**

Ключевые слова: гамма–пуассоновская модель; регрессия Кокса; гибридная байесовская сеть доверия; неточность измерений; последние эпизоды

I. Введение

Современные требования к анализу рисков организаций предполагают всесторонний учет и мониторинг разнообразных аспектов деятельности, в том числе и человеческих факторов, которые в ряде областей являются ключевыми [1]. Например, именно в результате действий человека происходит все большее число утечек данных в области кибербезопасности [2]. При этом уязвимости персонала активно используются злоумышленниками при реализации *социоинженерных атак* [3, 4].

Профиль уязвимостей пользователя опирается на разнородные данные о человеке, его психологические особенности [5] и склонность к рискообразующему поведению [6]. Последняя характеристика неявно отражена в паттерне реализации эпизодов на временной оси. Однако человек и его поведение представляют сложный для наблюдения объект, и потому зачастую при оценке риска приходится обращаться к информации, получаемой в результате самоотчетов (в том числе и об эпизодах поведения), которая является неполной и неточной. Использование подобной информации в системах управления риском опирается на математические модели [7].

Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН № FFZF-2022-0003

Таким образом, возникает задача оценки кумулятивной характеристики эпизодического поведения индивида, которая отражает временной ряд эпизодов рискообразующего поведения с учетом возможных внешних факторов, по неполным и неточным данным об эпизодах поведения. В целом, можно говорить о моделировании целого паттерна эпизодического поведения, включающего несколько взаимосвязанных объектов, которые могут наблюдаться с некоторой неточностью. Существуют два подхода: классический, в основе которого лежит регрессия Кокса [8, 9], и байесовский, который опирается на вероятностные графические модели с непрерывными переменными в узлах [6].

II. Обзор литературы

Подходы к учету эпистемической неточности данных в моделях линейной регрессии включают в себя как применения мягких вычислений, таких как теория нечетких множеств [10] и теория неопределенности [11], так и классические способы моделирования неопределенности: интервалы [12]. Однако применение подобных методов ограничено для особых типов регрессий, как модель Кокса, которая служит для анализа процессов повторяющихся событий.

Байесовские сети доверия (БСД) заточены под моделирование неопределенности различных типов. Некоторые формализации моделей БСД (в том числе с непрерывными переменными в узлах) допускают пропагацию интервальных свидетельств [13, 14]. Этот подход является классическим в задаче моделирования эпистемической неопределенности [16]. Кроме того, был разработан формализм нечетких байесовских сетей доверия [15].

Целью работы является определение устойчивости имеющихся подходов к оценке интенсивности поведения по неполным данным к ошибкам и неточностям данных об эпизодах поведения индивидов. Теоретическая значимость исследования заключается в подходе к сравнению двух методов оценивания искомой характеристики. Практическая значимость работы заключается в установлении границ применимости этих методов в приложениях. Новизна исследования состоит в сравнительном анализе преимуществ и недостатков двух подходов в условиях неопределенности.

III. Методология исследования

A. *Гамма-пуассоновская модель поведения и подходы к построению оценок интенсивности поведения*

Итак, в качестве основного объекта, отражающего поведение индивида, рассматривается последовательность эпизодов на временной оси. Пусть наблюдаются n индивидов, каждый предоставляет информацию о d_i эпизодах поведения, произошедших в определенный промежуток времени. При этом число d_i часто невелико в силу искажения припоминания. Математической моделью для процесса повторяющихся событий является точечный случайный процесс N(t), который характеризуется функцией интенсивности λ(t) [17]. Отметим, что гамма-пуассоновская модель предполагает гетерогенность популяции, т. е. что интенсивность поведения варьируется между пользователями.

Классическим подходом к оценке функции интенсивности такого процесса является регрессия Кокса. Чтобы учесть гетерогенность популяции, рассматривается регрессионная модель с со случайным параметром frailty в виде [17]:

$$\lambda(t) = z\,\lambda_0(t)\,exp\,(\beta x),$$

где x представляет матрицу неслучайных ковариант или факторов, которые оказывают влияние на процесс поведения, β – вектор коэффициентов регрессии, $\lambda_0(t)$ есть базовая функция интенсивности и z – случайная переменная, отражающая гетерогенность выборки, т. е. не учтенные в матрице X индивидуальные различия в интенсивности поведения. Подгонка такой регрессионной модели осуществляется с помощью модели пропорциональных рисков Кокса, в которую добавлен случайный фактор, frailty [17].

Модель байесовской сети доверия, отражающая особенности гамма-пуассоновской модели поведения, представлена в [7], и формализована с использованием копулы-лозы для включения непрерывных переменных в работе [6]. Подход к обработке неопределенности данных об интервалах на основе латентных переменных в классической модели БСД для гамма-пуассоновской модели представлен в [18]. Краеугольным камнем моделей, основанных на классических БСД, является дискретизация по своей сути непрерывных переменных – длин интервалов между эпизодами поведения. Чтобы обойти это ограничение, возможно использование *гибридных байесовских сетей доверия* [6]. Существуют различные подходы к учету непрерывных переменных в рамках формализма БСД. В работе использован подход на основе базисных функций, который позволяет приближать распределения непрерывных переменных с помощью набора усеченных экспонент или полиномов [19], что допускает проведение байесовского вывода.

БСД предназначены для анализа неопределенности, и наиболее частым способом учета неточности измерений является включение в модель латентных переменных, которые отражают истинное значение.

B. *Алгоритм имитационного моделирования*

Исследование опирается на имитационное моделирование. Алгоритм состоит из следующих шагов.

1. Создание выборки 1000 наблюдений для 5 возможных параметров гамма-распределения вероятности: (0.1, 1); (0.3,1); (1, 1). Выбор таких значений параметров обусловлен, с одной стороны, интерпретируемостью среднего значения частоты поведения (один эпизод в семь дней, один эпизод в три дня, один эпизод в два дня, один эпизод в день и два эпизода в день соответственно), а с другой стороны простотой вычислительных операций.

2. Для каждого значения интенсивности сгенерировать выборку из трех экспоненциально распределенных значений, соответствующих длинам интервалов между эпизодами поведения. Взять подвыборку значений, отвечающих длинам интервалов менее 365 дней.

3. Для каждого значения интенсивности на основе значений длин интервалов, полученных на шаге 2, создать зашумленные значения длин интервалов с ошибкой. В рамках работы моделью такой невязки выступает равномерное распределение на отрезке от -0.2 длины интервала до 0.2 длины интервала между последовательными эпизодами поведения.

4. При помощи специализированных пакетов вычислить оценку параметра frailty и соответствующий доверительный интервал (ДИ) в модели регрессии Кокса для исходных и зашумленных данных.

5. При помощи специализированных пакетов оценить параметры гибридной байесовскую сеть доверия.

6. Создать на основе исходной выборки 100 подвыборок меньшего объема (100 наблюдений), каждую из которых использовать для проведения байесовского вывода для получения оценки интенсивности поведения по данным о трех последних эпизодах поведения.

7. Построить эмпирические доверительные интервалы для интенсивности поведения на основе исходных и зашумленных данных, полученных на шаге 6.

IV. Результаты исследования

Для достижения цели исследования было проведено имитационное моделирование процесса эпизодического поведения индивида согласно разработанному алгоритму. Статистическое моделирование проводилось в среде обработки данных R, пакет emfrail [20] (подгонка регрессии Кокса) и пакет MoTBFs [19] (байесовский вывод для гибридных байесовских сетей доверия).

В табл. I представлены оценки, полученные при помощи классического подхода регрессии Кокса со случайным фактором (frailty).

ТАБЛИЦА I. Оценки исходного параметра интенсивности поведения по исходным и зашумленным данным с использованием подхода регрессии Кокса

Исходное значение параметра	**Оценка параметра, метод регрессии и 95% ДИ для исходных данных**	**Оценка параметра и 95% ДИ для зашумленных данных**
0.3	1.144 (0.946; 1.409)	1.126 (0.933, 1.385)
0.1	0.257 (0.231; 0.288)	0.257 (0.231,0.288)

В табл. II представлены значения параметров гамма-распределения, полученные с помощью гибридной байесовской сети доверия для исходных и для зашумленных данных.

ТАБЛИЦА II. ОЦЕНКИ ИСХОДНОГО ПАРАМЕТРА ИНТЕНСИВНОСТИ ПОВЕДЕНИЯ ПО ИСХОДНЫМ И ЗАШУМЛЕННЫМ ДАННЫМ С ИСПОЛЬЗОВАНИЕМ ПОДХОДА ГИБРИДНОЙ БАЙЕСОВСКОЙ СЕТИ ДОВЕРИЯ

Исходное значение параметра	**Среднее значение оценки параметра и 95% ДИ для исходных данных**	**Среднее значение оценки параметра и 95% ДИ для зашумленных данных**
1	0.752 (0.426, 1.076)	0.742 (0.464,1.125)
0.3	0.260 (0.131, 0.424)	0.248 (0.123, 0.388)
0.1	0.509 (0.176, 0.893)	0.486 (0.100, 1.01)

V. ОБСУЖДЕНИЕ

Результаты вычислительных экспериментов показывают, что оценка искомого параметра гамма–распределения вероятности при помощи модели регрессии превышает исходное значение, однако небольшая ошибка при измерении длины интервала между эпизодами поведения (в пределах +- 0.2 от истинной длины интервала) не приводит к значимому изменению получаемой оценки, особенно для малых значений исходной интенсивности поведения.

Оценивание искомого параметра гамма–распределения интенсивности поведения также возможно по слегка зашумленным данным, доверительные интервалы значительно перекрываются. Интересно отметить работу этого метода для низких значений параметра интенсивности: хотя среднее значение оценок достаточно велико, практически в пять раз выше исходного значения параметра, нижняя граница доверительного интервала близка к нему.

Среди ограничений исследования можно упомянуть небольшое число рассматриваемых моделей случайного шума и моделей эпизодического поведения, отраженных в значениях гамма-распределения интенсивности. Кроме того, поставленная задача рассмотрена с вероятностной точки зрения, тогда как подход на основе нечеткой логики является многообещающим при обработке данных с высокой степенью неопределенности и лингвистических переменных.

Предполагается, что дальнейшая работа будет направлена на преодоление этих ограничений, планируется рассмотреть различные модели зашумленности исходных данных и определить степень устойчивости подходов к оцениванию характеристик процесса поведения на основе сверхкороткого ряда наблюдений. В качестве ошибки можно рассмотреть и другие распределения: треугольное и нормальное. Планируется также исследование нечеткостной структуры модели эпизодического поведения индивидов.

VI. ЗАКЛЮЧЕНИЕ

Работа посвящена исследованию устойчивости существующих моделей оценивания характеристик поведения по ограниченным данным об эпизодах к умеренным ошибкам измерения длин интервалов между последовательными эпизодами. Такая ситуация часто возникает при работе с данными, полученными в результате самоотчетов респондентов. В работе представлен не только обзор существующих подходов к учету подобной неопределенности, но и проведены численные эксперименты, в ходе которых установлено, что точность регрессионных моделей для оценки частоты поведения невысока, в то время как подход на основе гибридные байесовских сетей доверия позволяет получать более точные оценки, особенно для высоких значений параметра интенсивности поведения.

СПИСОК ЛИТЕРАТУРЫ

[1] Zarei E., Yazdi M., Abbassi R., Khan F. A hybrid model for human factor analysis in process accidents: FBN-HFACS // Journal of loss prevention in the process industries. 2019. Т. 57. С. 142-155.

[2] Maalem Lahcen R. A., Caulkins B., Mohapatra R., Kumar M. Review and insight on the behavioral aspects of cybersecurity //Cybersecurity. 2020. Т. 3. №. 1. С. 1-18.

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

[4] Khlobystova A., Abramov M., Tulupyev A. An Approach to Building a Probabilistic Model of Spreading a Social Engineering Attack between Two Users //CEUR Workshop Proceedings. 2021. С. 53-58.

[5] Тулупьева Т. В., Абрамов М. В., Тулупьев А. Л. Модель социального влияния в анализе социоинженерных атак //Управленческое консультирование. 2021. №. 8 (152). С. 97-107.

[6] 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 //Proceedings of the Sixth International Scientific Conference “Intelligent Information Technologies for Industry”(IITI’22). Cham : Springer International Publishing, 2022. С. 465-474.

[7] Суворова А.В. Гибридные модели оценки параметров социально-значимого поведения по сверхмалой неполной совокупности наблюдений // Информатика и автоматизация. 2013. Т. 24. С. 116-134.

[8] Пащенко А. Е., Тулупьев А. Л., Николенко С. И. Статистическая оценка вероятности заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения //Труды СПИИРАН. 2006. Т. 2. №. 3. С. 257-268.

[9] Stoliarova V. F., Tulupyev A. L., Cox regression in the problem of risky behavior parameter estimation based on the last episodes’ data, St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 14 (4) (2021) 202–217. DOI: https://doi.org/10.18721/JPM.14415

[10] Chukhrova N., Johannssen A. Fuzzy regression analysis: systematic review and bibliography //Applied Soft Computing. 2019. Т. 84. С. 105708.

[11] Lio W., Liu B. Uncertain maximum likelihood estimation with application to uncertain regression analysis //Soft Computing. 2020. Т. 24. С. 9351-9360.

[12] Tretiak K., Schollmeyer G., Ferson S. Neural network model for imprecise regression with interval dependent variables //Neural Networks. 2023. Т. 161. С. 550-564.

[13] Koller D., Friedman N. Probabilistic graphical models: principles and techniques. MIT press, 2009.

[14] Joe H., Kurowicka D. (ed.). Dependence modeling: vine copula handbook. World Scientific, 2011.

[15] Pan Y., Zhang L., Li Z., Ding L. Improved fuzzy Bayesian network-based risk analysis with interval-valued fuzzy sets and D–S evidence theory //IEEE Transactions on Fuzzy Systems. 2019. Т. 28. №. 9. С. 2063-2077.

[16] Sahlin U., Helle I., Perepolkin D. “This Is What We Don't Know”: Treating epistemic uncertainty in bayesian networks for risk assessment //Integrated Environmental Assessment and Management. 2021. Т. 17. №. 1. С. 221-232.

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

[18] Toropova A.V., Tulupyeva T.V. Approbation of the Behavior Rate Model with Hidden Variables Based on Respondents' Data on Recent Instagram Posts //2021 XXIV International Conference on Soft Computing and Measurements (SCM). IEEE, 2021. С. 43-45.

[19] Pérez-Bernabé I., Maldonado A.D., Nielsen T.D., Salmerón A. Hybrid Bayesian Networks Using Mixtures of Truncated Basis Functions //R Journal. 2020. Т. 12. №. 2. С. 321-341.

[20] Balan T.A., Putter H. frailtyEM: An R package for estimating semiparametric shared frailty models //Journal of Statistical Software. 2019. Т. 90. С. 1-29.