УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2021 Том 21 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2021 Vol. 21 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2021-21-5-727-737
УДК 004.891 311.2

Машинное обучение байесовской сети доверия как инструмента оценки интенсивности процесса по данным из социальной сети

Александра Витальевна Торопова[1], Максим Викторович Абрамов[2], Татьяна Валентиновна Тулупьева[3]✉

[1,2] Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация
[2,3] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация
[3] СЗИУ РАНХиГС, Санкт-Петербург, 199178, Российская Федерация

[1] alexandra.toropova@gmail.com, https://orcid.org/0000-0001-7311-6192
[2] mva@dscs.pro, https://orcid.org/0000-0002-5476-3025
[3] tvt@dscs.pro✉, https://orcid.org/0000-0003-3630-7971

Аннотация

Предмет исследования. Рассмотрена задача оценки интенсивности протекания процессов, у которых математической моделью выступают стохастические процессы. Процессы состоят из серии последовательных эпизодов с известным классом распределений длины временного интервала между ними. Ранее был предложен подход, в котором входными данными выступали сведения о значении величины интервала между последним эпизодом и концом исследуемого периода, что могло привести к неточным результатам. Интервал отличался от промежутков между последовательными эпизодами, и поэтому его представление и обработка требует специальных подходов. Для повышения точности результатов оценки интенсивности процесса разработана новая модель. Модель основана на байесовской сети доверия и содержит узлы, которые соответствуют интервалам между последними эпизодами процесса, минимальным и максимальным интервалами между эпизодами, с помощью корректного учета на этапе обучения модели значений интервала между последним эпизодом и концом исследуемого периода. **Метод.** Предложена байесовская сеть доверия со случайным элементом, для построения интервала между окончанием периода исследования и последним эпизодом процесса за исследуемый период. На этапе обучения данные об этом интервале могут быть доступны. Для моделирования использовано программирование в системе R и пакет bnlearn, который обеспечивает работу с байесовскими сетями доверия. **Основные результаты.** Предложен новый подход к оценке интенсивности процесса на основе байесовской сети доверия, сформированной методами машинного обучения. Он позволяет повысить точность результатов посредством корректного учета величины интервала между последним эпизодом и окончанием исследуемого периода посредством применения особой схемы в машинном обучении байесовской сети, которая включает «гипотетический» эпизод после конца исследуемого периода. Для апробации предложенного подхода использованы данные о 5608 пользователях социальной сети Instagram на основании публикаций постов за 2020 год и первого поста за 2021 год. 70 % выборки использовано для обучения модели и 30 % для сравнения значений интенсивности постинга, предсказанных моделью с известными значениями. **Практическая значимость.** Полученные результаты могут применяться в различных сферах науки, где требуется оценка интенсивности процесса в условиях дефицита информации, когда весь процесс наблюдается ограниченное время. Получение таких оценок – важная задача в медицине, эпидемиологии, социологии и др. Подход показал хорошие результаты на сопоставлении теоретической модели и результатов обучения по данным из социальной сети, что создает основу для автоматизации получения оценок интенсивности процесса.

Ключевые слова

интенсивность процесса, оценка интенсивности, байесовские сети доверия, эпизоды процесса, стохастический процесс

Благодарности

Работа выполнена в рамках проекта по государственному заданию Санкт-Петербургского Федерального исследовательского центра Российской академии наук № 0073-2019-0003; при финансовой поддержке РФФИ: проект № 19-37-90120; проект № 20-07-00839.

Ссылка для цитирования: Торопова А.В., Абрамов М.В., Тулупьева Т.В. Машинное обучение байесовской сети доверия как инструмента оценки интенсивности процесса по данным из социальной сети // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 5. С. 727–737. doi: 10.17586/2226-1494-2021-21-5-727-737

Machine learning of the Bayesian belief network as a tool for evaluating the process frequency on social network data

Aleksandra V. Toropova[1], Maxim V. Abramov[2], Tatiana V. Tulupyeva[3]✉

[1,2] Saint Petersburg State University, Saint Petersburg, 199034, Russian Federation
[2,3] St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation
[3] Russian Presidential Academy of National Economy and Public Administration, Saint Petersburg, 199178, Russian Federation

[1] alexandra.toropova@gmail.com, https://orcid.org/0000-0001-7311-6192
[2] mva@dscs.pro, https://orcid.org/0000-0002-5476-3025
[3] tvt@dscs.pro✉, https://orcid.org/0000-0003-3630-7971

Abstract

The paper considers the problem of evaluating frequency of the processes whose mathematical model is stochastic processes consisting of a series of sequential episodes with a known class of distributions of the length of the time interval between them. In the previously proposed approach, the input data included information about the value of the interval between the last episode and the end of the study period, which could lead to inaccurate results. This interval differs from the intervals between successive episodes, and hence its presentation and processing require approaches that take this feature into account. Accuracy of the estimation results for process frequency was improved by developing a new model based on the Bayesian confidence network that includes nodes corresponding to the intervals between the last episodes of the process, the minimum and maximum intervals between episodes, by correctly accounting for the values of the interval between the last episode and the end of the study period at the model training stage. The authors propose a Bayesian belief network that includes a random element characterizing the interval between the end of the study period and the last episode of the process during the study period; data on this interval can be available at the training stage. They used R programming and the bnlearn package to model the Bayesian belief network. A new approach to the estimation of process frequency based on the Bayesian belief network generated by machine learning methods is proposed. It allows increasing the accuracy of the results by correctly considering the value of the interval between the last episode and the end of the period under study using a special scheme in the machine learning Bayesian belief network which includes a "hypothetical" episode after the end of the study period. To test the proposed approach, data was collected on 5608 Instagram users, which included the time of posting for the year 2020 and the time of publishing the first post for the year 2021. 70 % of the sample was used to train the model, and 30 % was used to compare the posting frequency values predicted by the model with known values. The results can be used in various fields of science, where it is necessary to estimate a process frequency under information deficit, when the whole process is observed for no more than some limited time. Obtaining such estimates is often an important issue in medicine, epidemiology, sociology, etc. The approach shows good results on the comparison of the theoretical model and the results of learning from the social network data, which can automate the obtaining of process frequency estimates.

Keywords

process frequency, frequency estimation, Bayesian belief networks, process episodes, stochastic process

Acknowledgements

The work was carried out within the framework of the project under the state assignment of St. Petersburg Federal Research Center of the Russian Academy of Sciences No. 0073-2019-0003 and with financial support from the Russian Foundation for Basic Research, projects No. 19-37-90120, No. 18-01-00626 and No. 20-07-00839.

For citation: Toropova A.V., Abramov M.V., Tulupyeva T.V. Machine learning of the Bayesian belief network as a tool for evaluating the process frequency on social network data. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 5, pp. 727–737 (in Russian). doi: 10.17586/2226-1494-2021-21-5-727-737

Введение

Исследования различных процессов, их интенсивности, моделирования, влияния условий на их протекание и др., свойственны многим областям науки, в том числе таким как информационная безопасность, управляющие системы, робототехника, эпидемиология и другие [1–4]. Особенности конкретных предметных областей и причины, по которым в этих областях встает задача оценить интенсивность тех или иных процессов в условиях информационного дефицита рассмотрены далее в разделе «Результаты и их обсуждение». При реализации возникают задачи автоматизации оценки интенсивности процесса (частоты его эпизодов) в условиях дефицита информации: по неточным, неполным данным, когда доступны лишь сведения о величине интервалов между несколькими последними эпизодами и сведения о предельных значениях интервалов между эпизодами [1, 3, 5, 6].

Интенсивность (частота) — одна из основных характеристик процесса, которая содержательно может быть интерпретирована как среднее число эпизодов,

случившихся в течение определенного периода времени [7]. Оценка интенсивности процесса позволяет предсказать его поведение, обнаружить закономерности, но зачастую получить оценку невозможно по ряду причин: сбор данных занимает много времени, дорого стоит и трудноисполним [8–10].

Известны подходы к оценке интенсивности процесса на основе информации о значениях величин интервалов между последними эпизодами процесса и рекордных значениях интервалов [1, 3, 5, 6]. В данных подходах при обучении модели использовано в том числе значение интервала между последним эпизодом и концом исследуемого периода. Поскольку момент окончания исследуемого периода не является эпизодом процесса, предлагаемая ранее модель могла не полностью раскрывать свой потенциал. В настоящей работе предложена математическая модель оценки интенсивности процесса по данным о значениях величин между последними эпизодами процесса, а также предельными значениями величин интервалов, отличающаяся от ранее рассмотренных тем, что при ее обучении корректно учитывается значение интервала между последним эпизодом процесса и концом исследуемого периода. Модель допускает последующую автоматизацию за счет применения байесовских сетей доверия (БСД) [1, 11], доступного программного обеспечения по их представлению, обработке, машинному обучению.

Модель связей между параметрами процесса, а также наблюдаемыми и ненаблюдаемыми реализациями его проявлений как математического объекта, представляет собой БСД. Этот инструмент выбран по наличию преимуществ: развитые и свободно доступные пакеты программ; возможность совмещать экспертную информацию и имеющиеся данные; при рассмотрении более сложных отношений и включений в модель новой информации (узлов) имеется возможность без труда перестроить модель [12–17]. Апробация предлагаемой модели выполнена на основе данных, извлеченных из социальной сети Instagram[1].

Теоретическая и практическая значимость построения модели заключается в возможности изучить аналитическими либо численными методами характеристики ее результатов и применить их в ряде прикладных областей: информационной безопасности, информационно-управляющих системах, эпидемиологии, социологии и других для оценки интенсивности процессов.

Постановка задачи

Существует класс задач, в которых необходимо моделировать процесс по неполным данным. Одним из примеров такой задачи может служить задача оценки интенсивности эпизодов поведения человека по информации о значениях интервалов между: предпредпоследним и предпоследним эпизодами; предпоследним и последним эпизодами; максимальным и минимальным значениями интервалов между эпизодами. Результаты данной задачи рассмотрены в работах [1, 5, 6, 18], но их точность пока еще недостаточно высока.

Цель исследования — повышение точности результатов оценки интенсивности процесса за счет разработки новой модели, основанной на БСД, включающей в себя узлы, соответствующие интервалам между последними эпизодами процесса, минимальным и максимальным интервалами между эпизодами, посредством корректного учета на этапе обучения модели значений интервала между последним эпизодом и концом исследуемого периода. Математической моделью исследуемого процесса выступает стохастический процесс, распадающийся на отдельные эпизоды.

Результатом, полученным в работе, служит модель, которая принимает на входе: непустое множество значений интервалов между последними тремя эпизодами процесса (при их наличии); значения интервала между завершением периода исследования и последним эпизодом; рекордные интервалы (минимальный и максимальный) между эпизодами за исследуемый период. На выходе модель дает оценку интенсивности процесса.

Релевантные работы

Определение интенсивности или частоты процесса — задача, встречающаяся во многих областях, так как это один из важнейших показателей процесса и, зная его, можно сделать выводы о процессе в целом. Например, в здравоохранении и медицине существуют задачи, для решения которых важно знать частоту заболеваний, интенсивность приема лекарств, интенсивность физических нагрузок и других процессов, связанных со здоровьем человека, чтобы принимать меры для эффективного лечения, профилактики и предупреждения заболеваний. Во многих медицинских исследованиях, в том числе занимающихся частотой заболеваний, большая проблема — сбор данных. В [19] были собраны ретроспективные данные о пациентах, зарегистрированных в амбулаторных отделениях пяти медицинских учреждений за три года (90 119 пациентов) для исследования частоты появления обсессивно-компульсивного расстройства, но из собранных данных, подходящими для исследования, оказалось только 65,8 %, а частота обсессивно-компульсивного расстройства была определена как 4,1 %. В [20] по статистике поступивших коммерческих страховых претензий от пациентов определена примерная интенсивность появления болезни Лайма в США за 2010–2018 гг. (примерно 476 000 пациентов ежегодно). В исследовании отмечено существенное занижение оценки частоты болезни Лайма в данных, полученных из медицинских источников (примерно 35 000 случаев ежегодно), поэтому авторы обращают внимание на то, что в имеющихся медицинских данных имеются заметные пропуски. На основании этих данных отметим, что в области медицины задача оценки частоты заболеваний или других процессов по неполным данным актуальна и важна, но зачастую трудновыполнима. Использование модели оценки интенсивности процесса на данных о последних эпизодах, которой посвящена настоящая работа, может частично восполнить нехватку медицинских данных.

[1] Instagram [Электронный ресурс]. Режим доступа: https://www.instagram.com (дата обращения: 10.08.2021).

В социологии интенсивность процесса также является объектом исследования. В [21] представлены результаты исследования, задачей которого было оценить зависимости между частотой приготовления обеда и ужина в домашних условиях со страной проживания респондента, его полом и восприятием субъективного благополучия. Были использованы данные всемирного опроса Гэллапа в 2018–2019 гг. (145 417 респондентов), собранные по 142 странам с помощью телефонных и личных интервью. Чтобы оценить частоту приготовления пиши респондентам задавался следующий вопрос: «Если подумать о последних 7 днях, в течение скольких из них Вы лично готовили [обед или ужин] у себя дома?». Для верификации ответов респондентов, либо моделирования частоты приготовления пищи в течение более длительного срока может быть полезен подход, предложенный в настоящей работе.

Еще один способ измерения интенсивности поведенческих процессов — прямой вопрос о частоте. В [22] исследована частота насильственных действий детей по отношению к родителям: с помощью опросника были собраны данные о 1543 студентах испанских институтов. На вопрос: «Как часто вы предпринимали психологическое/физическое/финансовое насилие по отношению к матери/отцу», — можно было ответить по пятибалльной шкале: 0 (никогда не было), 1 (было один раз), 2 (было два или три раза), 3 (было четыре или пять раз) и 4 (было шесть раз или больше). Использование таких шкал может исказить реальные данные об интенсивности процесса, поэтому использование предлагаемого в работе решения может способствовать получению более точных оценок о частоте насилия по отношению к родителям.

В экономике и менеджменте частота различных процессов имеет большое влияние. В [23] показано, что частота путешествий человека определяет такой параметр, как желание заплатить (willingness to pay, WTP), значение WTP достигало пика при оплате размещения 6 раз за 2 года), данные о частоте путешествий получены с помощью прямого вопроса: «Как часто вы оплачивали жилье за последние 24 месяца?» Вопросы о последних эпизодах съема жилья и оценка частоты путешествий по ответам на них могут быть использованы для получения более полных сведений.

БСД была использована для построения модели оценки интенсивности в пуассоновской и гамма-пуассоновской моделях случайных процессов. В [1, 5] предложенная модель принимает на вход данные о величине интервалов между последними тремя эпизодами процесса, а также минимальной и максимальной величинами интервалов. Для построения БСД значения интервалов между эпизодами дискретизируются.

В [1, 6, 18] были предприняты попытки усовершенствования модели стохастических связей между параметром и характеристиками процесса для более точных предсказаний: в [1] предложена модель со структурой, обученной автоматически; в [6] в модель добавлены узлы, показывающие согласованность данных; в [18] в модель добавлены скрытые переменные, характеризующие реальные интервалы между эпизодами (с учетом того, что полученные от респондента данные могут содержать ошибки и неточности). Включение во входные данные модели значений величины интервала между последним эпизодом и концом исследуемого периода, как интервала между эпизодами, может стать причиной неточностей. В частности, подход к обработке систематической ошибки, являющейся следствием неявного предположения о том, что конец исследования является эпизодом процесса, описан в [24]. Попытка построить функцию максимального правдоподобия в описанных обстоятельствах предпринята в [25], но ее результат оказался крайне неудачен для анализа и последующей реализации в программном комплексе.

Предпосылки и описание исследования

Формулировка задачи оценки интенсивности процесса по последним эпизодам близка к задачам из области временных рядов [26], но инструментарий временных рядов не подходит, так как для их моделирования необходимо большее количество данных об эпизодах. Даже для моделирования короткого временного ряда необходимо от 40 наблюдений, кроме того, данное моделирование характеризуется невысоким качеством. Для определения оценки интенсивности процесса может быть использован регрессионный анализ [27]. Преимущество байесовских сетей по сравнению с этим методом — высокая интерпретируемость и понятность моделей. Таким образом, в настоящей работе предложено использовать БСД, как и в работе [1].

Рассмотрим модель оценки интенсивности процесса, предложенную в [1]. Модель является БСД, структура которой представлена на рис. 1. Вершины характеризуют случайные элементы, входящие в модель; ребра — причинно-следственные связи между ними. Отметим, что t_{01}, t_{12}, t_{23}, $t_{\min}$, $t_{\max}$, λ — случайные величины, но при переходе к дискретизации за счет разбиения множества возможных значений на интервалы, используются случайные элементы, тесно связанные с исходными случайными величинами.

Случайная величина λ характеризует интенсивность процесса, n определяет количество эпизодов процесса за исследуемый период; t_{01}, t_{12} и t_{23} — значения величин интервалов между: последним эпизодом и концом исследуемого периода, последним и предпоследним эпизодами, предпредпоследним и предпоследним эпи-

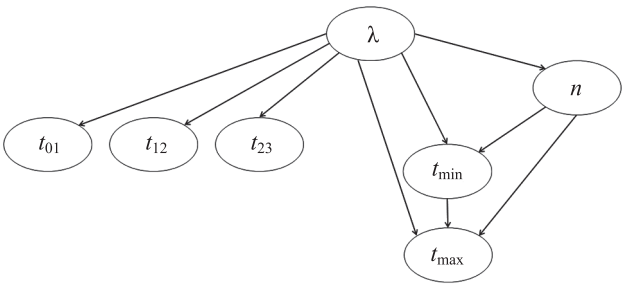

Рис. 1. Модель стохастических связей между параметром и характеристиками процесса как математического объекта [1]

Fig. 1. Model of stochastic relationships between a parameter and the characteristics of a process as a mathematical object [1]

зодами соответственно; $t_{\min}$ и $t_{\max}$ — сведения о минимальном и максимальном значениях величин интервалов между эпизодами за исследуемый период.

Кроме структуры, для определения БСД необходимо задать тензоры условной вероятности (P), характеризующие зависимости между узлами сети

$$P = \{P(t_{01}|\lambda), P(t_{12}|\lambda), P(t_{23}|\lambda), P(t_{\min}|n, \lambda), P(t_{\max}|n, \lambda, t_{\min}), P(n|\lambda), P(\lambda)\},$$

где $P(t_{ij}|\lambda)$ — условная вероятность того или иного значения t_{ij} при фиксированном значении λ; $P(t_{\min}|n, \lambda)$ — условная вероятность значений $t_{\min}$ при фиксированных значениях λ и n; $P(t_{\max}|n, \lambda, t_{\min})$ — условная вероятность значений $t_{\max}$ при фиксированных значениях λ, n и $t_{\min}$; $P(n|\lambda)$ — условная вероятность значений n при фиксированном значении λ; $P(\lambda)$ — вероятность того или иного значения λ.

Это можно сделать, обучив модель на статистических данных: сначала области значений величин, входящих в модель (временные интервалы, интенсивность и количество эпизодов) разбиваются на дискретные интервалы, после этого вычисляются условные вероятности для каждой пары переменных, соединенных ребром (см. далее раздел «Обучение байесовской сети доверия»). Тензоры условной вероятности можно представить в виде так называемых таблиц условной вероятности, т. е. многомерных таблиц из условных вероятностей вершин при условии различных означиваний их родителей.

Рассмотрим математическую модель процесса как стохастический процесс, распадающийся на отдельные эпизоды на временной оси, ограничиваясь промежутком между началом периода исследования и точкой, соответствующей окончанию исследуемого периода. Включим в рассмотрение промежуток между точкой окончания периода исследования и моментом следующего эпизода процесса. Для этого добавим в модель стохастических связей между характеристиками процесса вершину t_1^* (рис. 2), соответствующую интервалу между последним эпизодом за исследуемый период и следующим эпизодом, произошедшим после окончания исследуемого периода. Таким образом, в предложенной модели по сравнению с исходной производится более корректный учет значений интервала между последним эпизодом и окончанием исследуемого периода.

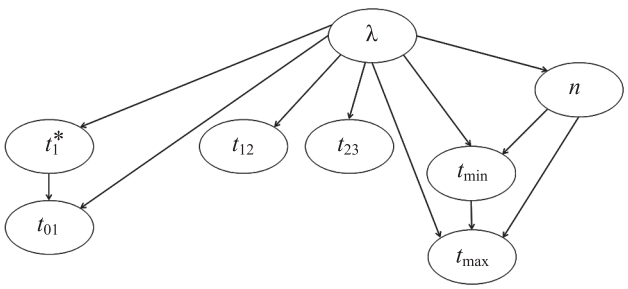

Рис. 2. Расширенная модель стохастических связей между параметром и характеристиками процесса как математического объекта

Fig. 2. Extended model of stochastic relationships between a parameter and the characteristics of a process as a mathematical object

Отметим, что при машинном обучении модели значение t_1^* доступно, оно присутствует в наборе данных, что позволяет учесть влияние соответствующей информации в строящейся БСД. Когда сеть построена, при ее практическом применении t_1^* уже не будет доступна. Например, при опросе об эпизодах поведения респондент не знает, когда в будущем будет иметь место следующий эпизод. Но это и не нужно; модель уже учитывает отношения между вовлеченными величинами, а оценки интенсивности процесса будут строится по доступным сведениям.

Для наглядности на рис. 3 представлен пример стохастического процесса: черными точками обозначены эпизоды процесса, прямоугольником — период исследования, точкой пересечения временной оси и прямоугольника справа (t_0) — момент окончания периода исследования, точками t_1, t_2 и t_3 отмечены последние три эпизода за исследуемый период, t^* — первый эпизод процесса, произошедший по окончании исследуемого периода.

Для обучения модели стохастических связей между параметром и характеристиками процесса (см. раздел «Обучение байесовской сети доверия») используем полный набор данных — для каждого отдельного процесса означены все вершины, входящие в модель. В случае отсутствия подходящего набора данных для изучаемого процесса можно использовать синтетические данные.

Порядок работы модели: для каждого случая означиваются вершины, соответствующие поступившим свидетельствам, рассчитывается апостериорная вероятность случайной величины — интенсивности процесса, и в качестве оценки интенсивности выдается интервал с наибольшей получившейся вероятностью [11].

Следующий этап исследования — апробация модели: предсказанные значения интенсивности процесса сравниваются с известными значениями интенсивности в тестовом наборе данных (см. раздел «Анализ работы модели»), и выполняется оценка, насколько удачным получилось предсказание.

Идею подхода к оценке можно проиллюстрировать на примере. Авторами собраны точные данные обо всех походах в магазин жителей некоего города за 2020 год. Изучим и вычислим интенсивность походов для каждого жителя взяв отношение количества походов к 366 дням (для 2020 года). Для каждого жителя известно время трех последних походов в магазин (при их наличии). Для части жителей также учтено количество походов, выполненных в 2021 году до начала исследования. На основании данных о 70 % жителей обучим модель, а на данных об оставшихся жителях

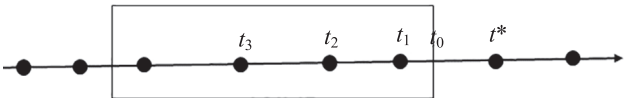

Рис. 3. Пример стохастического процесса

Fig. 3. Example of a stochastic process

проверим насколько близки оценки интенсивности модели к известным данным о походах этих жителей в магазин. Для получения результата необходимо, чтобы на наборе реальных данных удалось оценить, насколько удачно модель предсказывает значения интенсивности. Соответствующий подход и набор данных представлены в разделе «Данные» настоящей работы.

Для работы с байесовскими сетями доверия применен пакет bnlearn[1]. Анализ работы модели выполнен с помощью языка R[2]. В работе использовано округление до тысячных.

Описание модели

Предложена модель оценки интенсивности процесса на основе БСД со структурой, заданной экспертно. В отличие от предложенной в работе [1] в структуру новой модели включена вершина t_1^* (рис. 2), характеризующая интервал между последним эпизодом за исследуемый период и следующим эпизодом, произошедшим после окончания исследуемого периода. Отметим, что t_1^* использована только на этапе машинного обучения соответствующей БСД.

При построении БСД выполним разбиение области допустимых значений непрерывных величин на конечное число интервалов. В иллюстративных целях и для простоты понимания выбрана дискретизация областей возможных значений величин, представленная ниже; однако подход обобщен на произвольную дискретизацию. Для случайной величины λ, характеризующей интенсивность процесса, использована дискретизация вида: $\lambda^{(1)} = [0, 0{,}002)$, $\lambda^{(2)} = [0{,}002, 0{,}01)$, $\lambda^{(3)} = [0{,}01, 0{,}03)$, $\lambda^{(4)} = [0{,}03, 0{,}05)$, $\lambda^{(5)} = [0{,}05, 0{,}1)$, $\lambda^{(6)} = [0{,}1, 0{,}2)$, $\lambda^{(7)} = [0{,}2, 0{,}5)$, $\lambda^{(8)} = [0{,}5, 1)$, $\lambda^{(9)} = [1; \infty)$; для случайных величин $t_{i,j+1}$, t_1^*, $t_{\min}$ и $t_{\max}$, характеризующих длины интервалов между эпизодами, — $t^{(1)} = [0, 0{,}1)$, $t^{(2)} = [0{,}1, 0{,}5)$, $t^{(3)} = [0{,}5, 1)$, $t^{(4)} = [1, 7)$, $t^{(5)} = [7, 14)$, $t^{(6)} = [14, 30)$, $t^{(7)} = [30, 180)$, $t^{(8)} = [180, 365)$, $t^{(9)} = [365, \infty)$. В данном случае время измеряется в днях, а интенсивность процесса — как отношение числа эпизодов за исследуемый период к числу дней, составляющих этот период.

Данные

Для апробации модели необходимы реальные значения интенсивности процесса, для их сравнения с предсказанными значениями.

Данные для обучения и тестирования модели получены из социальной сети Instagram. Сайт этой социальной сети входит в число десяти наиболее популярных в России и один из наиболее посещаемых в мире[3]. В качестве исследуемого процесса рассмотрим публикацию пользователями постов в Instagram. Отметим, что речь идет именно о постах, публикации в «сторис» (временные публикации) не учитываются. На данный момент публикация постов — основной вид активности в Instagram. Пост представляет собой фотографию, видео или несколько фотографий в виде галереи, время публикации поста фиксируется в формате unixtime[4] (т. е. время представляется как прошедшее количество секунд с 1 января 1970 года до публикации). Отследим какое число постов опубликовано за интересующий нас период для вычисления интенсивности процесса.

На языке C# написана программа для сбора данных из сети Instagram. С помощью языка запросов GraphQL[5] получены сведения о постах пользователя в формате JSON. В качестве исследуемого периода выбран 2019 год.

По ID пользователя с помощью программы определим время публикации трех его последних постов за 2019 год, время публикации первого поста с 1 января 2020 года, минимальный и максимальный интервалы между публикациями за 2019 год и общее число постов за год. Вычислим значение интенсивности публикаций постов в Instagram, как отношение числа постов за исследуемый период к числу дней в этом периоде.

Случайным образом выберем 5608 пользователей с открытыми аккаунтами, из которых можно извлечь информацию о времени публикации постов. Для каждого пользователя рассчитываем интервалы между последними тремя публикациями за 2019 год, интервал между последней публикацией за 2019 год и первой в 2020 году, а также интервал между публикацией за 2019 год и концом исследуемого периода (1 января 2020 года). Каждый из интервалов отнесем к одному из интервалов дискретизации $t^{(1)}$–$t^{(9)}$. Если между постами прошло 15 дней, то значение относится к интервалу $t^{(6)} = [14, 30)$. Отметим, что в случаях, когда пользователи публикуют посты редко, либо вообще не имеют постов, значения параметров модели будут попадать в полуинтервал $t^{(9)} = [365, \infty)$.

Таким образом для каждого пользователя получены значения параметров t_1^*, t_{01}, t_{12}, t_{23}, $t_{\min}$, $t_{\max}$, n и рассчитана λ.

Итоговый набор составлен из информации о времени публикаций постов 5608 пользователей. Важно, что для каждого пользователя получено исходное значение интенсивности за год. Это делает возможным его сравнение с конечной оценкой.

Разделим выборку на две части, для обучения модели используем 3926 записей (70 % от выборки) и 1682 записей (30 % от выборки) — для последующей оценки и тестирования модели. На рис. 4 показано распределение значений интенсивности для тестирования модели.

[1] Bnlearn [Электронный ресурс]. Режим доступа: https://www.bnlearn.com (дата обращения: 10.08.2021).

[2] R Core Team. R Foundation for Statistical Computing. Vienna, Austria [Электронный ресурс]. Режим доступа: https://www.R-project.org (дата обращения: 10.08.2021).

[3] SimilarWeb [Электронный ресурс]. Режим доступа: https://www.similarweb.com/website/instagram.com (дата обращения: 10.08.2021).

[4] The Open Group Base Specifications [Электронный ресурс]. Режим доступа: https://pubs.opengroup.org/onlinepubs/9699919799/xrat/V4_xbd_chap04.html (дата обращения: 10.08.2021).

[5] GraphQL [Электронный ресурс]. Режим доступа: https://graphql.org (дата обращения: 10.08.2021).

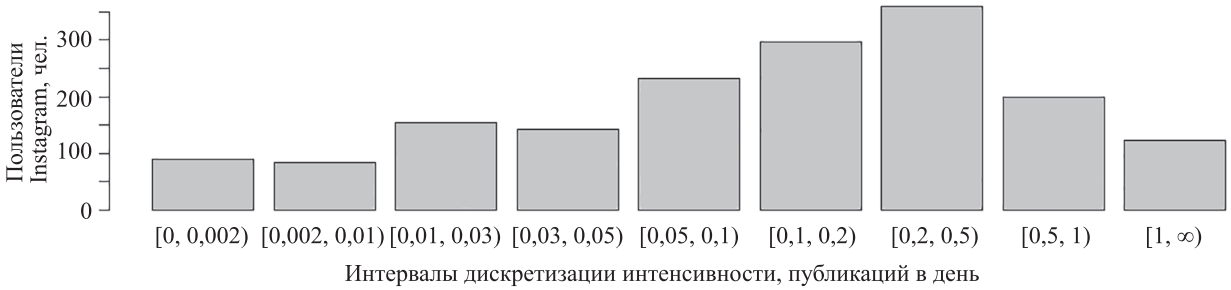

Рис. 4. Распределение интенсивности для тестирования модели
Fig. 4. Behavior rate distribution for the testing model

Обучение байесовской сети доверия

После получения данных, приведем их к виду, подходящему для машинного обучения параметров модели, вычислим значения интервалов между эпизодами в днях, полученное значение отнесем к одному из интервалов дискретизации t (см. раздел «Описание модели»). Если интервала между эпизодами не было (например, при отсутствии эпизода после исследуемого периода, соответственно нет интервала между ним и последним эпизодом за исследуемый период), определим его как принадлежащий к большему интервалу дискретизации временных интервалов, т. е. к $t^{(9)} = [365, \infty)$. Значение интенсивности публикации постов за год также отнесем к одному из интервалов дискретизации λ. На полученных данных выполним машинное обучение БСД. Для каждой пары переменных, соединенных ребром, вычислим условные вероятности. В данном исследовании используется наиболее часто применяемый алгоритм для полного набора данных (для каждого случая у нас означены все вершины) — метод максимального правдоподобия. Построим таблицы условных вероятностей: для вершин, у которых нет родителей, определяется частота появления их значений в данных; у вершин, имеющих родителей, определяется частота появления значений в данных при заданных значениях родителей. Сначала строится таблица условных вероятностей для величины λ, у вершины, соответствующей этой величине, нет родителей, таким образом получается одномерная таблица (табл. 1), ее значения — это отношения количества значений, относящихся к определенному интервалу, ко всем значениям.

Для вершин, у которых единственный родитель — вершина λ, получатся двумерные таблицы. Например, таблица условных вероятностей $P(t_{12}|\lambda)$ выглядит следующим образом (табл. 2). Сначала для t_{12} вычисляются количество значений, относящихся к определенному интервалу при $\lambda^{(1)}$, потом при $\lambda^{(2)}$, и так далее.

Обратим внимание на интервал $\lambda^{(1)} = [0, 0{,}002)$, к этому интервалу относятся объекты процесса, не имеющие эпизодов публикаций постов за исследуемый период (если произошел хотя бы один эпизод, то интенсивность равна 0,003 и не входит в этот интервал).

Таблица 1. Таблица вероятностей $P(\lambda)$
Table 1. Conditional Probabilities Table $P(\lambda)$

$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$
0,053	0,047	0,076	0,084	0,144	0,174	0,223	0,122	0,077

Таблица 2. Таблица условных вероятностей $P(t_{12}|\lambda)$
Table 2. Conditional Probabilities Table $P(t_{12}|\lambda)$

Интервал между эпизодами	Значение интенсивности								
	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$
$t^{(1)}$	0	0,043	0,084	0,076	0,071	0,067	0,073	0,058	0,146
$t^{(2)}$	0	0,005	0,007	0,021	0,034	0,015	0,055	0,104	0,286
$t^{(3)}$	0	0,005	0,034	0,052	0,042	0,069	0,106	0,221	0,282
$t^{(4)}$	0	0,108	0,191	0,238	0,315	0,420	0,572	0,568	0,272
$t^{(5)}$	0	0,032	0,104	0,119	0,147	0,205	0,115	0,035	0,013
$t^{(6)}$	0	0,070	0,151	0,192	0,207	0,146	0,065	0,010	0
$t^{(7)}$	0	0,296	0,403	0,302	0,181	0,077	0,014	0,002	0
$t^{(8)}$	0	0,059	0,027	0	0,002	0	0	0	0
$t^{(9)}$	1	0,382	0	0	0,002	0,001	0	0	0

Анализ работы модели

Основная задача построенной модели процесса — автоматизация оценки его интенсивности. Иначе говоря, необходимо автоматизировать оценку интенсивности процесса (частоты его эпизодов) в условиях дефицита информации: по неточным, неполным данным, когда доступны лишь сведения о величине интервалов между несколькими последними эпизодами, минимальном и максимальном значениях величин интервалов между эпизодами. При обозначенной выше дискретизации значений интенсивности эта задача представляет собой задачу классификации по девяти классам. Рассмотрим предсказания модели относительно интенсивности. В качестве исходных данных в модель загружаются сведения об интервалах между моментом окончания исследуемого периода и последним эпизодом, последним эпизодом и предпоследним эпизодом, предпоследним и предпредпоследним эпизодами за исследуемый период, а также данные о минимальном и максимальном интервалах – сведения, соответствующие вершинам t_{01}, t_{12}, t_{23}, $t_{\min}$, $t_{\max}$. На рис. 5 представлено распределение интенсивности, предсказанное моделью.

В табл. 3 представлена матрица ошибок (confusion matrix) значений интенсивности. Строки представляют собой исходные значения интенсивности, а столбцы — предсказанные моделью.

Точность (accuracy) — соотношение числа верно предсказанных значений, относительно всех значений, составила 0,549. Средняя точность (average accuracy) — доля правильно классифицированных значений в сумме матриц ошибок для каждого класса равна 0,899. Чтобы рассчитать среднюю точность, построим матрицы ошибок для каждого класса отдельно, вычислим сумму их диагоналей (значения, верно отнесенные к классу и верно не включенные в класс) и разделим на количество классов и общее число значений. Другие часто используемые характеристики precision, recall и f1-score равны 0,619, 0,589 и 0,604 соответственно. Заметим, что в большинстве случаев неверно предсказанные значения интенсивности оказались отнесены к интервалам, смежным к исходному значению, что не ведет к существенному для предметной области смещению оценки. Смежная точность (adjacent accuracy) — отношение правильно классифицированных и отнесенных к соседним классам значений к общему числу, равна 0,908 — величина данного показателя значительна и является важным достижением.

Результаты и их обсуждение

Отметим, что при оценке результатов качества модели стоит рассматривать именно среднюю точность, учитывающую количество классов при дискретизации переменной соответствующей интенсивности. Другие

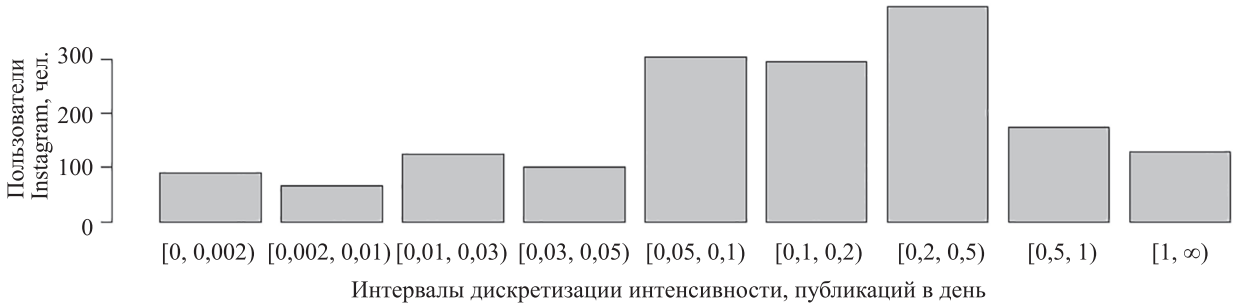

Рис. 5. Распределение интенсивности, предсказанное моделью

Fig. 5. Predicted behavior rate distribution

Таблица 3. Предсказание модели об интенсивности публикаций в Instagram

Table 3. Instagram posts behavior rate prediction

Исходное значение интенсивности	Предсказанное значение интенсивности								
	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$
$\lambda^{(1)}$	90	0	0	0	0	0	0	0	0
$\lambda^{(2)}$	0	63	16	1	0	1	0	0	0
$\lambda^{(3)}$	0	3	79	28	26	13	5	0	1
$\lambda^{(4)}$	0	0	23	38	59	16	7	0	0
$\lambda^{(5)}$	0	0	5	22	122	66	15	2	1
$\lambda^{(6)}$	0	0	2	8	74	126	80	5	1
$\lambda^{(7)}$	0	0	0	3	21	69	223	37	6
$\lambda^{(8)}$	0	0	0	0	2	3	58	100	37
$\lambda^{(9)}$	0	0	0	0	0	1	8	31	83

Таблица 4. Сравнение показателей качества моделей
Table 4. Comparison of common prediction quality metrics

Модель	Метрики качества моделей				
	точность	средняя точность	precision	recall	F1
Исходная [1]	0,541	0,897	0,610	0,582	0,596
С гипотетически «следующим» эпизодом	0,549	0,899	0,619	0,589	0,604

значения метрик моделей (точность, precision, recall и f1-score) оказались малы в сравнении со средней точностью, в том числе из-за довольно большого количества классов или интервалов дискретизации. При изучении различных процессов данный показатель может оказаться удобным для исследователей, так как можно задать дискретизацию интенсивности процесса, используя экспертные данные. В сравнении тех же данных с исходной моделью [1], представленная в данной работе модель имеет более высокие показатели (табл. 4).

Полученные результаты могут быть применены в различных сферах науки, где требуется оценка интенсивности процесса в условиях дефицита информации. Например, в контексте информационной безопасности в качестве процесса можно рассмотреть рискованное поведение пользователя информационной системы, заключающееся, допустим, в посещениях им непроверенных интернет-ресурсов, через которые распространяется вредоносное программное обеспечение, или открытие документов от неизвестных отправителей. Такое поведение потенциально может привести к компрометации системы. Для получения данных в этом случае можно использовать: метод дневника, когда респонденты ведут соответствующий дневник в течение некоторого времени, а затем передают экспертам; программное логирование; наблюдение и др. [28]. Обозначенные методы не всегда могут быть применены, поскольку требуют много времени для реализации и дороги. На основе анализа интенсивности такого процесса можно оценить вероятность компрометации информационной системы [3], эффективность мер по обеспечению защиты пользователей от социоинженерных атак [3, 5]. В рамках психологии как на процесс, можно смотреть на наблюдаемое поведение человека, чему посвящено целое направление – бихевиоризм или поведенческая психология [4, 29–32]. Подобных примеров множество, что свидетельствует об актуальности исследований в области интенсивности процессов, их моделирования, влияния условий на протекание [1–4].

Отметим, что благодаря наличию большого количества инструментов для работы с байесовскими сетями доверия, реализующих их функционал в программном виде (например, bnlearn), появляется возможность автоматизации получения оценки интенсивности процессов.

Заключение

В работе показано, как с помощью машинного обучения построить байесовскую сеть доверия, позволяющую по значениям величин интервалов времени: между окончанием периода исследования и последним, последним и предпоследним, предпоследним и предпредпоследним эпизодами, а также максимального и минимального интервалов между эпизодами за исследуемый период, получить оценку интенсивности процесса. На основе вычислительных экспериментов и статистической обработки их результатов установлено, что данная модель позволяет повысить точность оценки интенсивности случайного процесса. Отметим, что весь процесс доступен для наблюдения на ограниченное время, а модель обучена на данных, извлеченных из социальной сети Instagram. При анализе характеристик модели сопоставлены характеристики реальных данных об интенсивности процесса постинга пользователей и предсказания (оценки), полученные с помощью модели. Результаты, полученные в работе, создают основу для апробации модели на данных о других реальных процессах, например, о поведении. Стоит отметить, что при этом важно будет кроме сведений о последних эпизодах также получить сведения о фактической интенсивности протекания процесса. Результаты работы дают возможность автоматизации получения оценок интенсивности процессов с помощью описанной модели.

Литература

1. Суворова А.В., Тулупьев А.Л. Синтез структур байесовской сети доверия для оценки характеристик рискованного поведения // Информационно-управляющие системы. 2018. № 1. С. 116–122. https://doi.org/10.15217/issn1684-8853.2018.1.116
2. Conners E.E., West B.S., Roth A.M., Meckel-Parker K.G., Kwan M.-P., Magis-Rodriguez C., Staines-Orozco H., Clapp J.D., Brouwer K.C. Quantitative, qualitative and geospatial methods to characterize HIV risk environments // PLoS ONE. 2016. V. 11. N 5. P. e0155693 https://doi.org/10.1371/journal.pone.0155693
3. Абрамов М.В., Тулупьева Т.В., Тулупьев А.Л. Социоинженерные атаки: социальные сети и оценки защищенности пользователей. СПб.: ГУАП, 2018. 266 с.
4. Skinner B.F. Science and Human Behavior. Free Press, 1965. 461 p.

References

1. Suvorova A.V., Tulupyev A.L. Bayesian belief network structure synthesis for risky behavior rate estimation. *Informatsionno-Upravliaiushchie Sistemy*, 2018, no. 1, pp. 116–122. (in Russian). https://doi.org/10.15217/issn1684-8853.2018.1.116
2. Conners E.E., West B.S., Roth A.M., Meckel-Parker K.G., Kwan M.-P., Magis-Rodriguez C., Staines-Orozco H., Clapp J.D., Brouwer K.C. Quantitative, qualitative and geospatial methods to characterize HIV risk environments. *PLoS ONE*, 2016, vol. 11, no. 5, pp. e0155693 https://doi.org/10.1371/journal.pone.0155693
3. Abramov M.V., Tulupeva T.V., Tulupev A.L. *Social Engineering Attacks: Social Media and Users Security Estimates*. St. Petersburg, SUAI, 2018, 266 p. (in Russian)
4. Skinner B.F. *Science and Human Behavior*. Free Press, 1965, 461 p.

5. Суворова А.В. Моделирование социально-значимого поведения по сверхмалой неполной совокупности наблюдений // Информационно-измерительные и управляющие системы. 2013. Т. 11. № 9. С. 34–37.
6. Торопова А.В., Суворова А.В., Тулупьев А.Л. Диагностика согласованности в модели для оценивания интенсивности социально-значимого поведения // Нечеткие системы и мягкие вычисления. 2015. Т. 10. № 1. С. 93–107.
7. Friman P.C. Cooper, heron, and heward's applied behavior analysis (2nd edition): Checkered flag for students and professors, Yellow flag for the field // Journal of Applied Behavior Analysis. 2010. V. 43. N 1. P. 161–174. https://doi.org/10.1901/jaba.2010.43-161
8. Bolger N., Davis A., Rafaeli E. Diary methods: capturing life as it is lived // Annual Review of Psychology. 2003. V. 54. P. 579–616. https://doi.org/10.1146/annurev.psych.54.101601.145030
9. Graham C.A., Catania J.A., Brand R., Duong T., Canchola J.A. Recalling sexual behavior: A methodological analysis of memory recall bias via interview using the diary as the gold standard // Journal of Sex Research. 2003. V. 40. N 4. P. 325–332. https://doi.org/10.1080/00224490209552198
10. Kuleshov S., Zaytseva A., Aksenov A. Natural language search and associative-ontology matching algorithms based on graph representation of texts // Advances in Intelligent Systems and Computing. 2019. V. 1046. P. 285–294. https://doi.org/10.1007/978-3-030-30329-7_26
11. Тулупьев А.Л., Сироткин А.В., Николенко С.И. Байесовские сети доверия. СПб.: Изд-во Санкт-Петербургского ун-та, 2009. 399 с.
12. Dai J., Ren J., Du W. Decomposition-based Bayesian network structure learning algorithm using local topology information // Knowledge-Based Systems. 2020. V. 195. P. 105602. https://doi.org/10.1016/j.knosys.2020.105602
13. Bareinboim E., Pearl J. Causal inference and the data-fusion problem // Proceedings of the National Academy of Sciences of the United States of America. 2016. V. 113. N 27. P. 7345–7352. https://doi.org/10.1073/pnas.1510507113
14. Chen C., Zhang L., Tiong R.L.K. A novel learning cloud Bayesian network for risk measurement // Applied Soft Computing Journal. 2020. V. 87. P. 105947. https://doi.org/10.1016/j.asoc.2019.105947
15. Cobb B.R., Li L. Bayesian network model for quality control with categorical attribute data // Applied Soft Computing Journal. 2019. V. 84. P. 105746. https://doi.org/10.1016/j.asoc.2019.105746
16. He R., Tian J., Wu H. Structure learning in Bayesian networks of a moderate size by efficient sampling // Journal of Machine Learning Research. 2016. V. 17. P. 1–54.
17. Kabir G., Demissie G., Sadiq R., Tesfamariam S. Integrating failure prediction models for water mains: Bayesian belief network based data fusion // Knowledge-Based Systems. 2015. V. 85. P. 159–169. https://doi.org/10.1016/j.knosys.2015.05.002
18. Toropova A., Tulupyeva T. Synthesis and learning of socially significant behavior model with hidden variables // Advances in Intelligent Systems and Computing. 2019. V. 875. P. 76–84. https://doi.org/10.1007/978-3-030-01821-4_9
19. Jabeen S., Kausar R. Obsessive compulsive disorder: frequency and gender estimates // Pakistan Journal of Medical Sciences. 2020. V. 36. N 5. P. 1048–1052. https://doi.org/10.12669/pjms.36.5.1870
20. Kugeler K.J., Schwartz A.M., Delorey M.J., Mead P.S., Hinckley A.F. Estimating the frequency of lyme disease diagnoses, United States, 2010–2018 // Emerging Infectious Diseases. 2021. V. 27. N 2. P. 616–619. https://doi.org/10.3201/eid2702.202731
21. Wolfson J.A., Ishikawa Y., Hosokawa C., Janisch K., Massa J., Eisenberg D.M. Gender differences in global estimates of cooking frequency prior to COVID-19 // Appetite. 2021. V. 161. P. 105117. https://doi.org/10.1016/j.appet.2021.105117
22. Cano-Lozano M.C., León S.P., Contreras L. Child-to-Parent violence: examining the frequency and reasons in spanish youth // Family Relations. 2021. in press. https://doi.org/10.1111/fare.12567
23. Nieto-García M., Muñoz-Gallego P.A., Gonzalez-Benito Ó. The more the merrier? Understanding how travel frequency shapes willingness to pay // Cornell Hospitality Quarterly. 2020. V. 61. N 4. P. 401–415. https://doi.org/10.1177/1938965519899932
24. Зельтерман Д., Тулупьев А.Л., Суворова А.В., Пащенко А.Е., Мусина В.Ф., 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. № 1. С. 160–185. https://doi.org/10.15622/sp.16.6

5. Suvorova A.V. Socially significant behavior modeling on the base of super-short incomplete set of observations. *Information-measuring and Control Systems*, 2013, vol. 11, no. 9, pp. 34–37. (in Russian)
6. Toropova A.V., Suvorova A.V., Tulupyev A.L. Model for socially significant behavior rate estimate: consistency diagnostics. *Fuzzy Systems and Soft Computing*, 2015, vol. 10, no. 1, pp. 93–107. (in Russian)
7. Friman P.C. Cooper, heron, and heward's applied behavior analysis (2nd edition): Checkered flag for students and professors, Yellow flag for the field. *Journal of Applied Behavior Analysis*, 2010, vol. 43, no. 1, pp. 161–174. https://doi.org/10.1901/jaba.2010.43-161
8. Bolger N., Davis A., Rafaeli E. Diary methods: capturing life as it is lived. *Annual Review of Psychology*, 2003, vol. 54, pp. 579–616. https://doi.org/10.1146/annurev.psych.54.101601.145030
9. Graham C.A., Catania J.A., Brand R., Duong T., Canchola J.A. Recalling sexual behavior: A methodological analysis of memory recall bias via interview using the diary as the gold standard. *Journal of Sex Research*, 2003, vol. 40, no. 4, pp. 325–332. https://doi.org/10.1080/00224490209552198
10. Kuleshov S., Zaytseva A., Aksenov A. Natural language search and associative-ontology matching algorithms based on graph representation of texts. *Advances in Intelligent Systems and Computing*, 2019, vol. 1046, pp. 285–294. https://doi.org/10.1007/978-3-030-30329-7_26
11. Tulupev A.L., Sirotkin A.V., Nikolenko S.I. *Bayesian Belief Networks*. St. Petersburg, St Petersburg University Publ., 2009, 399 p. (in Russian)
12. Dai J., Ren J., Du W. Decomposition-based Bayesian network structure learning algorithm using local topology information. *Knowledge-Based Systems*, 2020, vol. 195, pp. 105602. https://doi.org/10.1016/j.knosys.2020.105602
13. Bareinboim E., Pearl J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences of the United States of America*, 2016, vol. 113, no. 27, pp. 7345–7352. https://doi.org/10.1073/pnas.1510507113
14. Chen C., Zhang L., Tiong R.L.K. A novel learning cloud Bayesian network for risk measurement. *Applied Soft Computing Journal*, 2020, vol. 87, pp. 105947. https://doi.org/10.1016/j.asoc.2019.105947
15. Cobb B.R., Li L. Bayesian network model for quality control with categorical attribute data. *Applied Soft Computing Journal*, 2019, vol. 84, pp. 105746. https://doi.org/10.1016/j.asoc.2019.105746
16. He R., Tian J., Wu H. Structure learning in Bayesian networks of a moderate size by efficient sampling. *Journal of Machine Learning Research*, 2016, vol. 17, pp. 1–54.
17. Kabir G., Demissie G., Sadiq R., Tesfamariam S. Integrating failure prediction models for water mains: Bayesian belief network based data fusion. *Knowledge-Based Systems*, 2015, vol. 85, pp. 159–169. https://doi.org/10.1016/j.knosys.2015.05.002
18. Toropova A., Tulupyeva T. Synthesis and learning of socially significant behavior model with hidden variables. *Advances in Intelligent Systems and Computing*, 2019, vol. 875, pp. 76–84. https://doi.org/10.1007/978-3-030-01821-4_9
19. Jabeen S., Kausar R. Obsessive compulsive disorder: frequency and gender estimates. *Pakistan Journal of Medical Sciences*, 2020, vol. 36, no. 5, pp. 1048–1052. https://doi.org/10.12669/pjms.36.5.1870
20. Kugeler K.J., Schwartz A.M., Delorey M.J., Mead P.S., Hinckley A.F. Estimating the frequency of lyme disease diagnoses, United States, 2010–2018. *Emerging Infectious Diseases*, 2021, vol. 27, no. 2, pp. 616–619. https://doi.org/10.3201/eid2702.202731
21. Wolfson J.A., Ishikawa Y., Hosokawa C., Janisch K., Massa J., Eisenberg D.M. Gender differences in global estimates of cooking frequency prior to COVID-19. *Appetite*, 2021, vol. 161, pp. 105117. https://doi.org/10.1016/j.appet.2021.105117
22. Cano-Lozano M.C., León S.P., Contreras L. Child-to-Parent violence: examining the frequency and reasons in spanish youth. *Family Relations*, 2021, in press. https://doi.org/10.1111/fare.12567
23. Nieto-García M., Muñoz-Gallego P.A., Gonzalez-Benito Ó. The more the merrier? Understanding how travel frequency shapes willingness to pay. *Cornell Hospitality Quarterly*, 2020, vol. 61, no. 4, pp. 401–415. https://doi.org/10.1177/1938965519899932
24. Zelterman D., Tulupyev A.L., Suvorova A.V., Paschenko A.E., Musina V.F., Tulupyeva T.V., Krasnoselskikh T.V., Grau L.E., Heimer R. Processing length bias of time intervals between the last episode and the interview in gamma-poisson models of behavior. *SPIIRAS Proceedings*, 2011, no. 1, pp. 160–185. (in Russian). https://doi.org/10.15622/sp.16.6

25. 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. С. 157–184. https://doi.org/10.15622/sp.23.9
26. Ярушкина Н.Г. Предиктивная аналитика на основе нечетких временных рядов // Интегрированные модели и мягкие вычисления в искусственном интеллекте (ИММВ-2021): Сборник научных трудов X Международной научно-технической конференции. В 2-х т. Т. 1. Коломна, 17–20 мая 2021 года. Смоленск: Универсум, 2021. С. 116–128.
27. Özkaya U., Yiğit E., Seyfi L., Öztürk S., Singh D. Comparative regression analysis for estimating resonant frequency of c-like patch antennas // Mathematical Problems in Engineering. 2021. V. 2021. P. 6903925. https://doi.org/10.1155/2021/6903925
28. Osipov V.Y., Vodyaho A.I., Zhukova N.A., Glebovsky P.A. Multilevel automatic synthesis of behavioral programs for smart devices // Proc. 2017 International Conference on Control, Artificial Intelligence, Robotics ans Optimization (ICCAIRO). 2017. P. 335–340. https://doi.org/10.1109/ICCAIRO.2017.68
29. Desmond N., Nagelkerke N., Lora W., Chipeta E., Sambo M., Kumwenda M., Corbett E.L., Taegtemeyer M., Seeley J., Lalloo D.G., Theobald S. Measuring sexual behaviour in Malawi: a triangulation of three data collection instruments // BMC Public Health. 2018. V. 18. N 1. P. 807. https://doi.org/10.1186/s12889-018-5717-x
30. Suvorova A., Belyakov A., Makhamatova A., Ustinov A., Levina O., Tulupyev A., Niccolai L., Rassokhin V., Heimer R. Comparison of satisfaction with care between two different models of HIV care delivery in St. Petersburg, Russia // AIDS Care. 2015. V. 27. N 10. P. 1309–1316. https://doi.org/10.1080/09540121.2015.1054337
31. Shane-Simpson C., Schwartz A.M., Abi-Habib R., Tohme P., Obeid R. I love my selfie! an investigation of overt and covert narcissism to understand selfie-posting behaviors within three geographic communities // Computers in Human Behavior. 2020. V. 104. P. 106158. https://doi.org/10.1016/j.chb.2019.106158
32. Chen S.X., Lam B.C.P., Hui B.P.H., Ng J.C.K., Mak W.W.S., Guan Y., Buchtel E.E., Tang W.C.S., Lau V.C.Y. Conceptualizing psychological processes in response to globalization: Components, antecedents, and consequences of global orientations // Journal of Personality and Social Psychology. 2016. V. 110. N 2. P. 302–331. https://doi.org/10.1037/a0039647

25. Stepanov D.V., Musina V.F., Suvorova A.V., Tulupyev A.L., Sirotkin A.V., Tulupyeva T.V. Risky behavior poisson model identification: heterogeneous arguments in likelihood. *SPIIRAS Proceedings*, 2012, no. 4, pp. 157–184. (in Russian). https://doi.org/10.15622/sp.23.9
26. Iarushkina N.G. Predicative analytics based on fuzzy time series. *"Integrated models and soft computing in artificial intelligence". Proceedings of the 10th International Scientific and Technical Conference. V. 1. Kolomna, May 17-20, 2021*. Smolensk, Universum Publ., 2021, pp. 116–128. (in Russian)
27. Özkaya U., Yiğit E., Seyfi L., Öztürk S., Singh D. Comparative regression analysis for estimating resonant frequency of c-like patch antennas. *Mathematical Problems in Engineering*, 2021, vol. 2021, pp. 6903925. https://doi.org/10.1155/2021/6903925
28. Osipov V.Y., Vodyaho A.I., Zhukova N.A., Glebovsky P.A. Multilevel automatic synthesis of behavioral programs for smart devices. *Proc. 2017 International Conference on Control, Artificial Intelligence, Robotics ans Optimization (ICCAIRO)*, 2017, pp. 335–340. https://doi.org/10.1109/ICCAIRO.2017.68
29. Desmond N., Nagelkerke N., Lora W., Chipeta E., Sambo M., Kumwenda M., Corbett E.L., Taegtemeyer M., Seeley J., Lalloo D.G., Theobald S. Measuring sexual behaviour in Malawi: a triangulation of three data collection instruments. *BMC Public Health*, 2018, vol. 18, no. 1, pp. 807. https://doi.org/10.1186/s12889-018-5717-x
30. Suvorova A., Belyakov A., Makhamatova A., Ustinov A., Levina O., Tulupyev A., Niccolai L., Rassokhin V., Heimer R. Comparison of satisfaction with care between two different models of HIV care delivery in St. Petersburg, Russia. *AIDS Care*, 2015, vol. 27, no. 10, pp. 1309–1316. https://doi.org/10.1080/09540121.2015.1054337
31. Shane-Simpson C., Schwartz A.M., Abi-Habib R., Tohme P., Obeid R. I love my selfie! an investigation of overt and covert narcissism to understand selfie-posting behaviors within three geographic communities. *Computers in Human Behavior*, 2020, vol. 104, pp. 106158. https://doi.org/10.1016/j.chb.2019.106158
32. Chen S.X., Lam B.C.P., Hui B.P.H., Ng J.C.K., Mak W.W.S., Guan Y., Buchtel E.E., Tang W.C.S., Lau V.C.Y. Conceptualizing psychological processes in response to globalization: Components, antecedents, and consequences of global orientations. *Journal of Personality and Social Psychology*, 2016, vol. 110, no. 2, pp. 302–331. https://doi.org/10.1037/a0039647

Авторы

Торопова Александра Витальевна — младший научный сотрудник, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, SC 57144053600, https://orcid.org/0000-0001-7311-6192, alexandra.toropova@gmail.com

Абрамов Максим Викторович — кандидат технических наук, руководитель лаборатории, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация; доцент, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, SC 56938320500, https://orcid.org/0000-0002-5476-3025, mva@dscs.pro

Тулупьева Татьяна Валентиновна — кандидат психологических наук, доцент, СЗИУ РАНХиГС, Санкт-Петербург, 199178, Российская Федерация; старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57144665900, https://orcid.org/0000-0003-3630-7971, tvt@dscs.pro

Authors

Aleksandra V. Toropova — Junior Researcher, Saint Petersburg State University, Saint Petersburg, 199034, Russian Federation, SC 57144053600, https://orcid.org/0000-0001-7311-6192, alexandra.toropova@gmail.com

Maxim V. Abramov — PhD, Head of Laboratory, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation; Associate Professor, Saint Petersburg State University, Saint Petersburg, 199034, Russian Federation, SC 56938320500, https://orcid.org/0000-0002-5476-3025, mva@dscs.pro

Tatiana V. Tulupyeva — PhD, Associate Professor, Russian Presidential Academy of National Economy and Public Administration, Saint Petersburg, 199178, Russian Federation; Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 57144665900, https://orcid.org/0000-0003-3630-7971, tvt@dscs.pro

Статья поступила в редакцию 06.08.2021
Одобрена после рецензирования 25.08.2021
Принята к печати 02.10.2021

Received 06.08.2021
Approved after reviewing 25.08.2021
Accepted 02.10.2021