Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Санкт-Петербургский государственный университет»

Федеральное государственное бюджетное учреждение науки
«Санкт-Петербургский Федеральный исследовательский центр
Российской академии наук»

На правах рукописи

ТОРОПОВА Александра Витальевна

МЕТОДЫ И АЛГОРИТМЫ ОБРАБОТКИ НЕОПРЕДЕЛЕННОСТИ ДАННЫХ ПРИ ОЦЕНИВАНИИ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА ПО ОГРАНИЧЕННОМУ ОБЪЕМУ ДОСТУПНЫХ НАБЛЮДЕНИЙ

Специальность 2.3.1 – Системный анализ, управление и обработка информации, статистика

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор физико-математических наук, профессор
ТУЛУПЬЕВ Александр Львович

Санкт-Петербург – 2022

Оглавление

ВВЕДЕНИЕ

Актуальность темы исследования. Исследования различных процессов, их интенсивности, влияния условий на их протекание свойственны многим социоориентированным областям знаний, таким как экономика, социология, информационная безопасность, эпидемиология, и др. В маркетинговых и экономических исследованиях анализируется поведение покупателей и пользователей каких-либо сервисов [7, 18, 33, 74, 139, 216]; в социологии — поведение человека при взаимодействии с другими, а также в социальных сетях [159, 182, 211, 255]; в информационной безопасности [26, 79] — поведение пользователей, характеризующее подверженность социоинженерным атакующим воздействиям. В области эпидемиологии [9, 22, 86, 205] поведение человека может быть связано с риском передачи неизлечимых заболеваний, таких как ВИЧ (вирус иммунодефицита человека) [108].

При этом специалистам в области здравоохранения часто необходимо оценивать такие параметры как кумулятивный риск распространения ВИЧ в популяции и индивидуальный риск заражения. Иными словами, актуальной является разработка решений оценивания ряда числовых параметров, связанных с моделью реализации эпизодов поведения человека на числовой оси, с целью использования в системах анализа риска и принятия решений.

Из-за невозможности организации прямого длительного наблюдения за поведением индивида, основным источником данных становятся самоотчеты респондентов об их поведении. Такие ответы часто даются на естественном языке, характеризуются нечеткостью, неточностью и неполнотой, что обусловливает актуальность задачи разработки инструментов косвенного оценивания параметров процесса, в частности его интенсивности. Кроме того, при анализе поведения человека может быть известна дополнительная информация, существенно влияющая на модель поведения, например, психологические свойства личности или социальные характеристики респондентов. Возможно также привлечение экспертной информации. Для корректного учета возникающей неопределенности

и возможной дополнительной информации используются математические модели поведения.

Одной из классических моделей такого эпизодического поведения выступает пуассоновский процесс [88, 119, 203, 254]. Для учета неопределенности могут использоваться подходы на основе нечетких временных рядов [78, 215], но данная методология к исследуемой задаче не может быть применена в связи с тем, что для моделирования нечетких временных рядов необходимо больше данных об эпизодах поведения (более 40 наблюдений) [78]. Также для решения задачи оценивания интенсивности пуассоновского процесса по ограниченному объему данных можно применять методы регрессионного анализа [224].

Однако такие подходы не позволяют в полной мере моделировать возникающую неопределенность. Так как данные о поведении подвержены намеренным и ненамеренным искажениям, актуальной является научная задача моделирования неопределенности, связанной с ответами респондентов. В диссертационном исследовании рассматривается неопределенность, связанная с искажением респондентами ответов на вопросы о последних эпизодах поведения, а также с ошибками при задании конца интервала наблюдения.

В диссертационном исследовании решается **научная задача** моделирования и обработки неопределенности, связанной с ответами респондентов.

Важность и значимость решаемой задачи обусловлены возможностью применения полученных результатов в различных социоориентированных областях, в которых требуется моделирование эпизодического поведения человека по ограниченному объему доступных наблюдений.

Степень разработанности темы. В работах А.Л. Тулупьева, С.И. Николенко, Т.В. Тулупьевой, А.Е. Пащенко и соавторов [21, 71, 72] для минимизации неточности информации, предоставляемой респондентом, было предложено использовать данные о нескольких последних последовательных эпизодах процесса и рекордных интервалах. Дальнейшее развитие этот подход получил в работах А.В. Суворовой, А.В. Сироткина, В.Ф. Столяровой [30–32, 224]. Для получения оценки интенсивности эпизодов поведения на временной оси с

учетом возникающей неопределенности получаемой информации об эпизодах было предложено использовать байесовские сети доверия (БСД). Однако предложенные модели не в полной мере учитывают неопределенность, связанную с получаемыми данными: не учитывается некорректность получаемых от респондента данных об эпизодах поведения, несогласованность таких данных или же некорректное задание момента окончания исследования.

Заделом для диссертационного исследования в том числе послужили работы Н.Г. Ярушкиной [78] и С.М. Ковалева [215], посвященные анализу сверхкоротких нечетких временных рядов; М.К. Кано-Лосано [106] и Дж.А. Вольфсон [247], отражающие подходы к оцениванию интенсивности определенных видов поведения; Р.А. Рехфельдта [202], описывающие методы сбора информации о поведении; С.А. Потрясаева, А.Л. Ронжина, Б.В. Соколова, С.В. Микони, Р.М. Юсупова и А.В. Смирнова, предлагающие подходы к моделированию сложных объектов и систем [13, 23, 24].

Цель исследования заключается в повышении качества классификации при оценивании интенсивности пуассоновского процесса как математической модели эпизодического поведения индивида за счет разработки методов и алгоритмов обработки неопределенности данных, предоставляемых респондентами.

Для достижения поставленной цели был сформулирован ряд **задач**:

1. Разработать метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания интенсивности пуассоновского процесса, основанной на байесовской сети доверия.

2. Разработать алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, для улучшения оценки с точки зрения показателей качества классификации.

3. Разработать метод и алгоритм обработки неопределенности задания конца исследуемого периода, при оценивании интенсивности пуассоновского

процесса по ограниченному объему доступных наблюдений, для улучшения оценки с точки зрения показателей качества классификации.

4. Разработать архитектуру и прототип комплекса программ, реализующих предложенные методы и алгоритмы, для их апробации, вычислительных экспериментов и решения практических задач.

Объектом исследования являются сверхкороткие, неполные и неточные данные о последовательных эпизодах и рекордных интервалах пуассоновского процесса.

Предметом исследования являются методы, алгоритмы и модели обработки неопределенности в задаче оценивания интенсивности пуассоновского процесса по сверхкоротким, неполным и неточным данным о последовательных эпизодах поведения и рекордных интервалах между эпизодами поведения за определенный промежуток времени.

Научная новизна диссертационной работы определяется тем, что:

1. Предложены новые метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса, отличающиеся применением расширенной, по отношению к используемой ранее, байесовской сети доверия с дополнительными узлами принятия решений, обеспечивающие возможность работы с данными необходимой степени согласованности.

2. Предложен новый алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, отличающийся использованием новых скрытых переменных в байесовской сети доверия, отвечающих истинным длинам интервалов, обеспечивающий повышение показателей качества классификации по сравнению с предложенными ранее подходами.

3. Предложены новые метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений на

основе байесовской сети доверия, отличающиеся внедрением вершины, характеризующей интервал между последним эпизодом пуассоновского процесса и эпизодом, произошедшим после окончания периода исследования, обеспечивающие повышение показателей качества классификации по сравнению с предложенными ранее подходами при наличии ретроспективных данных.

4. Разработаны архитектура и прототип комплекса программ для оценивания интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений, отличающиеся наличием инструментов обработки неопределенности данных, связанной с ответами респондентов, обеспечивающие работу с предложенными в диссертационной работе алгоритмами.

Теоретическая и практическая значимость работы. В рамках диссертации были разработаны математические модели на основе байесовских сетей доверия, которые могут быть использованы для оценивания интенсивности поведения и адаптированы к использованию информации со значительной долей неопределенности. Такие модели являются новыми и представляют теоретическую значимость.

Разработанные методы и алгоритмы создают основу для обработки неопределенности и некорректности информации, полученной от респондентов, при построении оценок интенсивности эпизодического поведения человека по ограниченному объему доступных наблюдений. Кроме того, предложенные методы и алгоритмы обеспечивают возможность автоматизации решения задачи построения оценок и составляют основу для создания систем поддержки принятия решений в социоориентированных областях знаний.

Методология исследования. Задача обработки неопределенности данных при оценивании числовых параметров пуассоновского процесса по ограниченному объему доступных наблюдений является мультидисциплинарной. Методология системного анализа позволяет представить рассматриваемую задачу как общую задачу разработки специального математического и алгоритмического обеспечения системы принятия решений и обработки информации в условиях информационного дефицита в различных социоориентированных областях.

Методы исследования. В аналитической части работы используются методы теории вероятностей для построения математической модели поведения (пуассоновский процесс) и методы байесовской статистики и машинного обучения для обработки возникающей неопределенности, в частности теория байесовских сетей доверия. Для получения данных были использованы методы поиска и сбора информации из социальных сетей, также были использованы методы синтеза данных и имитационное моделирование зашумленных эпизодов пуассоновского процесса со случайной интенсивностью.

Положениями, выносимыми на защиту, являются

1. Метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанной на байесовской сети доверия.

2. Алгоритм обработки некорректности информации об эпизодах поведения, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида.

3. Метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений.

4. Архитектура и прототип комплекса программ, реализующие разработанные методы и алгоритмы.

Соответствие диссертации научной специальности. Представленные результаты соответствуют специальности 2.3.1 — «Системный анализ, управление и обработка информации, статистика».

Успешная апробация результатов предложенной диссертации на научных конференциях различного уровня, в том числе международных и российских, публикации в рецензируемых изданиях, согласованность результатов, качественный анализ тематики и корректное использование приведенных моделей и математических методов обусловили высокую **степень достоверности** полученных в исследовании результатов.

Апробация результатов. Основные научные мероприятия, в ходе которых были представлены и обсуждались результаты предлагаемого диссертационного исследования:

1) XVIII–XXI, XXIII, XXIV, XXV Международная конференция по мягким вычислениям и измерениям (SCM), г. Санкт-Петербург, 2015–2018, 2020, 2021, 2022 гг.

2) XVIII Национальная конференция по искусственному интеллекту КИИ-2020, г. Москва, 2020 г.

3) VI Международная научно-практическая конференция ICIT–2020, г. Саратов, 2020 г.

4) I, III International Scientific Conference «Intelligent Information Technologies for Industry» (IITI), Sochi, 2016, 2018.

5) Конференция «Информационные технологии в управлении» (ИТУ), г. Санкт-Петербург, 2016, 2018 гг.

6) X–XI Санкт-Петербургская межрегиональная конференция «Информационная безопасность регионов России» (ИБРР). Санкт-Петербург, 2017, 2019 гг.

7) The 11th Conference of the European Society for Fuzzy Logic and Technology organized jointly with the IQSA Workshop on Quantum Structures, Prague, 2019.

8) IV Международная летняя школа-семинар по искусственному интеллекту для студентов, аспирантов, молодых ученых и специалистов «Интеллектуальные системы и технологии: современное состояние и перспективы» ISYT–2017, г. Санкт-Петербург, 2017 г.

9) VII– VIII всероссийская научно-практическая конференция «Нечеткие системы, мягкие вычисления и интеллектуальные технологии» (НСМВИТ), г. Санкт-Петербург, 2017 г., г. Смоленск, 2020 г.

10) III Всероссийская Поспеловская конференция с международным участием «Гибридные и синергетические интеллектуальные системы», г. Светлогорск, 2016 г.

11) Всероссийская научная конференция по проблемам информатики (СПИСОК), г. Санкт-Петербург, 2016, 2017 гг.

12) Научная сессия НИЯУ МИФИ-2015, г. Москва, 2015 г.

13) VIII Международная научно-техническая конференция «Интегрированные модели и мягкие вычисления в искусственном интеллекте», г. Коломна, 2015 г.

14) III Международная школа-семинар по искусственному интеллекту для студентов, аспирантов и молодых ученых «Интеллектуальные системы и технологии: современное состояние и перспективы», г. Тверь, 2015 г.

Полученные в диссертации результаты, являются частью научно-исследовательских проектов, поддержанных следующими грантами РФФИ:

1) «Машинное обучение и структурные особенности байесовской сети доверия со скрытыми переменными как модели социально-значимого поведения» № 19-37-90120, 2019–2021. (грант «Аспиранты»).

2) «Методы идентификации параметров социальных процессов по неполной информации на основе вероятностных графических моделей» № 16-31-00373, 2016–2017.

3) «Комбинированный логико-вероятностный графический подход к представлению и обработке систем знаний с неопределенностью: алгебраические байесовские сети и родственные модели» № 15-01-09001, 2015–2017.

Полученные результаты были использованы при проведении исследовательских работ СПб ФИЦ РАН, в учебном процессе СЗИУ РАНХиГС, а также при разработке подходящего для клиента режима физических нагрузок в ООО «Хоум Фитнес», получены соответствующие акты внедрения.

Публикации. Результаты диссертационного исследования нашли отражение в научных работах соискателя. По теме диссертации было сделано 42 публикации и научных работ, к ним приравненных. В это число входят:

- 1 монография;
- 12 публикаций в изданиях, индексируемых Scopus/WoS;

- 4 статьи в изданиях из «Перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук»;
- 23 доклада и тезиса на научных конференциях (из которых 13 единоличных);
- 1 публикация в рецензируемом журнале;
- получено 1 свидетельство о регистрации программ для ЭВМ (Роспатент).

Перечень публикаций соискателя по теме диссертации представлен в приложении Б. Журналы, входящие в перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, в которых были опубликованы статьи автора:

- Журнал «Компьютерные инструменты в образовании» [34];
- Журнал «Нечеткие системы и мягкие вычисления» [28];
- Журнал «Информатика и автоматизация» (ранее «Труды СПИИРАН») [43];
- Журнал «Научно-технический вестник информационных технологий, механики и оптики» [48].

А.В. Торопова выступала в качестве соавтора в некоторых научных публикациях.

Личный вклад А.В. Тороповой охарактеризован следующим образом. В [55] описан метод диагностики согласованности данных респондентов, основанный на расширении модели оценивания интенсивности пуассоновского процесса, проведено исследование расширенной модели. В [50–53, 56] проведено исследование модели оценивания интенсивности пуассоновского процесса, расширенной узлами согласованности на различных данных. В [49, 231, 232, 235] предложен метод обработки возможной некорректности информации, полученной

от респондентов, при оценивании интенсивности пуассоновского процесса, основанный на модели оценивания интенсивности процесса со скрытыми переменными, учитывающей возможную некорректность данных респондентов, проведен ее анализ, рассмотрена ее работа на различных данных, кроме этого в [235] рассмотрена модель оценивания интенсивности пуассоновского процесса со скрытыми переменными с обученной структурой. В [54] рассмотрены подходы к обработке «зашумленных» данных, проведены вычислительные эксперименты. В [58, 60, 233, 234, 236, 237] предложен метод обработки длины интервала между последним эпизодом процесса и окончанием периода исследования при оценивании интенсивности пуассоновского процесса, включающий построение модели оценивания интенсивности процесса с гипотетически «следующим» эпизодом поведения, проведены вычислительные эксперименты, собраны данные для их проведения, кроме того в [60] рассмотрена модель оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом со структурой обученной на данных, в [234] проведено сравнение предложенной модели с исходной моделью оценки интенсивности поведения, в [236, 237] собраны данные из социальной сети ВКонтакте, проведено обучение и тестирования моделей на собранных данных, проведен анализ полученных результатов, в [61] разработан опросный инструментарий для сбора данных о последних эпизодах постинга в Instagram*, собраны и обработаны данные, проведено тестирование модели на собранных данных, в [28] собраны данные, полученные в результате изучения информации в социальной сети ВКонтакте и проведены эксперименты по обучению и апробации моделей, в [48] предложена модель, собраны данные и проведены обучение и апробация модели.

Структура и объем диссертации. Диссертация включает в себя введение, четыре главы, заключение, список сокращений, список литературы (255 источников), списки иллюстративного материала и таблиц, приложения. Общий

* Организация, запрещенная на территории РФ и признанная экстремистской.

объем диссертации — 184 страницы, включая 35 таблиц, 52 рисунка и 3 приложения.

В **первой главе** обоснованы актуальность диссертационного исследования, использование байесовских сетей доверия для задачи оценивания интенсивности пуассоновского процесса, выступающего моделью эпизодического поведения. Проведен обзор сфер применения байесовских сетей доверия и инструментария для работы с ними. Рассмотрены возможные подходы к обучению байесовских сетей доверия.

Вторая глава посвящена описанию теоретических результатов, послуживших предпосылками для данного диссертационного исследования и создающих основу для решения поставленных задач. Описаны основы теории байесовской сети доверия. Представлен разработанный ранее подход к оцениванию интенсивности пуассоновского процесса как модели эпизодического поведения индивида по сверхкороткому набору наблюдений с использованием байесовской сети доверия. Также описываются используемые метрики качества моделей и структура исследования. Данная глава не включает результаты данного диссертационного исследования, а предназначена для введения единой системы обозначений и описания предложенного ранее подхода к моделированию пуассоновских процессов, в развитии которого заключается суть этой работы.

Третья глава представляет собой описание полученных соискателем теоретических результатов. Предложены метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанной на байесовской сети доверия; алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида; метод и алгоритм обработки длины интервала между последним эпизодом процесса и концом исследуемого периода, при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений. Рассмотрена модель оценивания интенсивности пуассоновского процесса, расширенная скрытыми

переменными, как истинными данными о последних эпизодах и рекордных интервалах процесса, со структурой, обученной на синтетических данных. Описаны модели оценивания интенсивности постинга в социальных сетях, расширенные за счет объективных данных о пользователе. Также рассмотрены возможные варианты дискретизации непрерывных величин, входящих в модели.

В четвертой главе описаны архитектура и прототип комплекса программ для работы с моделями оценивания интенсивности пуассоновского процесса, реализующего разработанные методы и алгоритмы, в который входят следующие модули: модуль для работы с инструментом оценивания согласованности информации о последних эпизодах и рекордных интервалах процесса; модуль для работы с моделью оценивания интенсивности пуассоновского процесса со скрытыми переменными; модуль для работы с моделью оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом. Также описаны собранные данные (синтетические и данные из социальных сетей) для апробации предложенных моделей и результаты этой апробации.

ГЛАВА 1. ЗАДАЧА ОБРАБОТКИ НЕОПРЕДЕЛЕННОСТИ ДАННЫХ ПРИ ОЦЕНИВАНИИ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА ПО ОГРАНИЧЕННОМУ ОБЪЕМУ ДОСТУПНЫХ НАБЛЮДЕНИЙ

В диссертационном исследовании решается задача обработки некоторых типов неопределенности данных при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений. В первой главе обоснована актуальность диссертационного исследования, даются примеры того, как при известном значении интенсивности пуассоновского процесса эпизодического поведения можно определить другие аспекты, связанные с поведением. Обосновано применение БСД, описаны сферы их использования. Проведен обзор инструментария для работы с БСД. Рассмотрены возможные подходы к их обучению.

Изложение материала основано на оригинальных авторских обзорах [34, 43].

1.1 АНАЛИЗ ПОДХОДОВ К ОБРАБОТКЕ НЕОПРЕДЕЛЕННОСТИ ДАННЫХ ПРИ ОЦЕНИВАНИИ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА ПО ОГРАНИЧЕННОМУ ОБЪЕМУ ДОСТУПНЫХ НАБЛЮДЕНИЙ

Во многих областях науки возникают задачи, связанные с моделированием и построением оценок параметров точечных случайных процессов. В случае невозможности или большой ресурсозатратности работы с реальными объектами или системами их математическое моделирование становится возможным решением [13, 23, 24]. В данном исследовании рассматривается однородный пуассоновский процесс, параметр которого отражает интенсивность реализации эпизодов на временной оси. При этом часто возникает ситуация, когда о процессе известен лишь ограниченный объем доступных наблюдений [88].

Под интенсивностью (частотой) пуассоновского процесса подразумевается отношение количества эпизодов пуассоновского процесса за период исследования к количеству временных единиц, составляющих этот период исследования [134]. Эпизод пуассоновского процесса — это активность, произошедшая в определенное время, у которой можно определить начало и конец. Таким образом, эпизод произошел, когда эта активность закончилась. Для примера можно рассмотреть некоторые процессы поведения: курение, употребление алкоголя, встречи с друзьями, посещение определенных веб-сайтов и т. д. Оценка интенсивности пуассоновского процесса, позволяет предсказать его дальнейшее поведение, обнаружить в нем какие-то закономерности, но зачастую получить эту оценку невозможно по ряду причин: сбор данных занимает много времени, дорого стоит и трудозатратен [100, 141].

Существует несколько вариантов для сбора данных. Наиболее надежным методом получения данных об интенсивности пуассоновского процесса, очевидно, является прямое наблюдение [179, 202]. При этом доступность прямого наблюдения ограничена: есть много ситуаций, в которых его может быть трудно или даже невозможно реализовать. Дороговизна, временные затраты, правовые аспекты, зачастую невозможность непосредственного наблюдения — это его основные недостатки. По этой причине исследователи стали применять в основном косвенные методы, то есть методы, позволяющие оценить характеристики процессов поведения на основе данных, предоставляемых респондентами.

Дневниковый метод [100, 173] подразумевает последовательную и подробную запись действий респондента в течении дня за определенный промежуток времени. После сбора этих данных эксперт подсчитывает количество эпизодов поведения определенного типа за исследуемый период. В настоящее время этот метод часто модифицируется с помощью специальных приложений для смартфонов, заменяющих физические дневники [207]. Недостаток такого подхода — низкая скорость получения данных.

Одним из наиболее популярных методов является самоотчет (self-report) [188, 202]: респонденты заполняют анкету или дают интервью о своем поведении. Отметим, что на задаваемые вопросы респонденты отвечают по памяти, и чем раньше произошли события, о которых их спрашивают, тем сложнее ответить на эти вопросы для респондентов, и тем больше они могут ошибаться.

Для самоотчета наиболее часто используются два метода. Первый — это прямой вопрос респондентам о том, сколько раз они что-то делали за определенный период. Данные ответы часто не соответствуют реальным значениям. Например, на вопрос о количестве яблок, съеденных в прошлом году, сложно ответить правильно. В [247] исследовались связи между тем, как часто респонденты готовят обеды и ужины дома, полом, страной проживания и их благополучием в связи с эпидемией Covid-19. Данные для исследования были собраны из телефонных и личных опросов (2018–2019 гг., 145 417 респондентов) резидентов 142 стран. Частота готовки определялась с помощью вопроса: «В течение последних семи дней сколько раз Вы готовили дома?». А в [190] было показано, как то, насколько человек часто путешествует влияет на его желание заплатить (willingness to pay, WTP), для сбора данных о частоте поездок был использован прямой вопрос: «Как часто Вы платили за проживание за последние два года?». Отметим, что ответы на такие вопросы часто не соответствуют реальным значениям [30].

Второй метод — это использование Лайкерт-шкал [155], содержащих качественные значения типа «всегда», «часто», «иногда», «редко», «никогда» и др. Например, в [106] Лайкерт-шкала была использована для оценивания частоты насилия со стороны детей к их родителям. Были опрошены 1543 студента из испанских ВУЗов. При опросе был использован вопрос: «Насколько часто вы совершали насильственные действия психологического, финансового или физического характера к матери или отцу?», в ответе нужно было указать одно из пяти значений: «такого никогда не было», «один раз», «два-три раза», «четыре-пять раз», «от шести раз и более». Легко составить вопросы и получить ответы, но из этих ответов сложно получится вывести числовую оценку интенсивности поведения, так как разные люди под одним и тем же качественным значением могут

понимать совершенно разные количественные значения и наоборот, например, для кого-то значение «5 раз в месяц» может означать «часто», а для кого-то — «редко». Исследователи обычно используют количественную интерпретацию этих шкал, но такие результаты не могут правильно оценить интенсивность поведения. Есть и другие факторы, снижающие эффективность использования Лайкерт-шкал: даже отображение шкалы Лайкерта в горизонтальном или вертикальном формате может повлиять на ответы респондентов [246].

С целью извлечения наиболее точной числовой информации об эпизодах поведения в работах [30–32] был предложен метод последних эпизодов и рекордных интервалов. Этот метод опирается на вопросы: «Когда Вы в последний / предпоследний / предпредпоследний раз совершали …?» и «Каков минимальный / максимальный временной интервал между …?». Ответы на такие вопросы обладают рядом преимуществ перед ответами на прямые вопросы: являются численными описаниями конкретной легко определяемой характеристики, момента реализации некоторого действия. Однако получаемая таким способом информация может быть подвержена некоторой неопределенности, а именно ответы респондентов могут быть некорректны, неточны или не согласованы, также момент интервью может быть некорректно задан.

Располагая данными об интенсивности пуассоновского процесса, реально не только оценить значимые аспекты поведения в текущем времени, но и делать достаточно уверенные прогнозы о будущих проявлениях таких аспектов. В [156, 157] предложен метод построения оценки вероятности успешной реализации социоинженерной атаки по результатам наблюдения за активностью пользователей социальных сетей и изменениям интенсивности их взаимодействия.

В [137] ответы респондентов об интенсивности и нагрузках их тренировок интерпретируются предложенными методами в прогноз количества тренировок в полугодовой и годовой перспективе.

Интенсивность участия отцов в чтении с детьми может быть учтена при прогнозировании способностей детей к чтению и математике [87, 120]. В [160] показано, что интенсивность взаимодействия между родителями и детьми влияет

на контролирующее или «теплое» отношение родителей к детям. В [166] было выявлено, что с увеличением интенсивности жестокого обращения в подростковом возрасте возрастает уровень насильственного оскорбительного поведения. В [89] показано, что по интенсивности запросов Google пользователей об изоляции можно сделать выводы о распространении COVID-19.

Подобных примеров множество, что свидетельствует об актуальности исследований в области моделирования эпизодического поведения человека, и, в частности, получении оценок его интенсивности.

1.2 КЛАССИФИКАЦИЯ МЕТОДОВ ОБРАБОТКИ НЕОПРЕДЕЛЕННОСТИ ДАННЫХ ПРИ ОЦЕНИВАНИИ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА ПО ОГРАНИЧЕННОМУ ОБЪЕМУ ДОСТУПНЫХ НАБЛЮДЕНИЙ

Есть ряд способов моделирования процесса поведения, однако при сборе данных обычно используются опросы респондентов, которые наиболее точно могут ответить лишь о нескольких последних эпизодах поведения и рекордных интервалах за какой-то определенный период [72]. Таким образом встает задача определения интенсивности пуассоновского процесса поведения по таким данным.

Классифицировать методы обработки неопределенности данных при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений можно следующим образом.

1. **Метод максимального правдоподобия.** В [27] была приведена формула, выведенная аналитически и основанная на методе максимального правдоподобия, для расчета интенсивности пуассоновского процесса при использовании данных о рекордных интервалах и последнем эпизоде поведения:

$$L_{1,\min,\max}(\lambda)=\sum_{n=2}^{\infty}\frac{e^{-\lambda I}(\lambda I)^n}{n!}\frac{\partial^3}{\partial x\partial y\partial z}\left[\begin{array}{l}F_{1,n}^{n}(x,y;\lambda)-\\-\int\limits_{\Omega(x,y)}V_{u,v}^{*(n-2)}(I-z)\,f_{1,n}^{n}(u,v;\lambda)dudv\end{array}\right]\Bigg|\tilde{t}_{01},\tilde{t}_{\min},\tilde{t}_{\max},$$ где

$f_{1,n}^{n}(u,v;\lambda)$ — совместная плотность распределения порядковых статистик с

номерами $1,n$, определяется по формуле: $f_{1,n}^{n}(u,v;\lambda)=n(n-1)(F(u)-F(v))^{n-2}f(u)f(v)$, где $F(x)$ — показательная функция распределения, $F_{1,n}^{n}(x,y;\lambda)$ — соответствующая функция распределения, $V_{u,v}^{*(n-2)}(I-z)$ — $(n-2)$-кратная свертка функций распределения $H_{u,v}(x;\lambda)=\dfrac{F(x)-F(u)}{F(u)-F(v)}=\dfrac{1-e^{\lambda(u-x)}}{1-e^{\lambda(u-v)}}$, множество $\Omega(x,y)=\{0\le u\le x,u<v\le y,x<y\}$, $\tilde{t}_{01}$ — длина между последним эпизодом и сбором данных, $\tilde{t}_{\min}$ и $\tilde{t}_{\max}$ — длины минимального и максимального интервалов за период исследования $[0;I]$ [27]. Данная формула для применения в практических задачах, требует дальнейшего исследования и разработки методов вычисления входящих в нее выражений, кроме того включение в нее дополнительных сведений об интервалах между эпизодами или дополнительных сведений о процессе будет очень сложно.

2. **Сверхкороткие временные ряды.** Формулировка исследуемой задачи получения оценки интенсивности пуассоновского процесса схожа с задачами из области сверхкоротких временных рядов [78, 215], однако в данном случае необходимо больше данных об эпизодах процессов, даже в коротком временном ряду необходимо от 40 наблюдений.

3. **Регрессионный анализ**. Регрессионный анализ используется для построения оценок интенсивности поведения по ограниченным данным об эпизодах поведения, однако этот метод не приспособлен к учету значительной неопределенности входных данных [224].

4. **Байесовские сети доверия**. В работах [30–32] БСД уже использовались для оценивания интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений, учитывая при этом неопределенность, связанную с гранулярностью ответов респондентов. Кроме того, БСД обладают необходимым функционалом для обработки неопределенности, возникающей при работе с неполными и неточными данными, возможностью совмещения экспертной и статистической информации, наглядного изменения структуры зависимостей, добавления дополнительных связей и

элементов и высокой интерпретируемостью. Также существует большое число программного обеспечения, реализующего методы работы с БСД (обучения и вероятностного вывода, см. раздел 1.6).

Таким образом, можно утверждать, что байесовские сети доверия являются оптимальным инструментом для работы с информацией, получаемой в результате опросов и интервью в исследуемой задаче.

1.3 БАЙЕСОВСКАЯ СЕТЬ ДОВЕРИЯ КАК ИНСТРУМЕНТ ДЛЯ ОБРАБОТКИ НЕОПРЕДЕЛЕННОСТИ

Как было указано использование БСД встречается во многих областях науки, где требуется обработка неопределенности, связанной с исходными данными, далее приведены примеры.

1.3.1 Область медицины и здравоохранения. В последнее время БСД стали популярны в этой области благодаря таким качествам, как возможность работы с неточными данными, а также удобное и легко интерпретируемое формальное представление неточных знаний за счет объединения инструментов искусственного интеллекта и статистических методов анализа данных [15, 73, 81].

При выборе инструментов для описания систем здравоохранения надо учитывать влияние различных факторов и взаимодействие между элементами, определяющих функционирование этих систем. Также важны понятность и интерпретируемость таких инструментов, возможность использования неточных данных. БСД обладают этими свойствами, благодаря чему довольно часто используются в указанной области [81].

В [191] предлагаются предметно-зависимые ограничения для того, чтобы учесть возможную недостаточность и некорректность экспертных медицинских данных при построении БСД.

Основные задачи медицины и здравоохранения заключаются в диагностике заболеваний, прогнозировании состояния обследуемых [85, 105, 163], выборе подходящего курса лечения [80, 177], обнаружение функциональных взаимодействий на клеточном уровне [151, 200] и др.

1.3.2 Экология. Байесовские сети доверия также являются довольно популярным инструментов в области экологии [102, 180]. Наиболее часто встречаются два варианта применения этого инструмента. Первый вариант — это понимание функционирования экологических систем и их связей. БСД применяются для выявления круга вовлеченных экологических процессов, при оценивании степени значимости воздействия процессов на результат, в выявлении их взаимодействия, полезного влияния прогнозирования переменных в наблюдаемых процессах. Другой вариант — это оценивание значений некоторых переменных [180].

В [133] с помощью БСД моделируются и прогнозируются проблемы в системе распределения питьевой воды. В [77] — оценивается экологическая ситуация в зоне влияния химически опасных объектов и вероятность опасных ситуаций, с ними связанных. В [164] — оценивается работа экосистемных служб управления озерным комплексом в Бельгии. В [238] создана модель для определения рисков угроз почвы, эта модель объединяет экспертные данные и результаты анализов почвы.

1.3.3 Функциональная безопасность — значительная составляющая в различных технических системах. В нее входят проектирование систем безопасности и предпринимаемые в них действия [10, 148]. В [181, 252] рассматривается функциональная безопасность энергосберегающих систем, в [25] — функциональная безопасность в бортовых системах управления, в [208] — в системе газового тракта реактивных двигателей. В [171] предложена модель системы производственных процессов на примере литейного завода, в [113] — системы доставки электропитания.

1.3.4 Экономика и риск-анализ. БСД также часто применяются в области экономики и риск анализа, где часто нужно комбинировать различные виды информации, как например статистические и экспертные данные [75], и, кроме того, принимать решения в условиях неопределенности [16, 245].

БСД нашли применение в моделировании рисков. Используя базу известных факторов риска, БСД разрабатывают возможные сценарии [101]. В [130]

представлена модель, разработанная для определения рисков и выявления источников. В [76, 117] проводятся моделирование и анализ операционных рисков. Это риски убытка, как следствие некорректных внутренних процессов, в том числе действия персонала, а также внешнего воздействия. БСД позволяют моделировать такие риски, если их возникновение зависит от времени [16] и при зависимости частоты возникновения риска и его последствий [186].

В [29] строятся модели информационных рисков (угроза информационной безопасности), в [165, 245] — надежность сложных систем, в [3, 11] оценивается надежность телекоммуникационных услуг. В [6, 253] строятся модели работы финансовых учреждений.

В [170] оценивается эффективность сотрудников и организационного здоровья компании. В [172] выявляются вероятные причины ошибок сотрудника и оценивается степень его надежности. В [183] с помощью БСД формулируются факторы внимания работника при контроле железнодорожного движения. В [1] оценивалась вероятность работы без ошибок оператора сложного технологического объекта.

В [158, 169] с помощью БСД планируется расписание проекта в условиях неопределенности.

В [127, 239] строятся модели риск-анализа для морского судоходства. В [19] предлагается подход к оцениванию навигационной безопасности.

В [96, 225] моделируются риски природных катаклизмов.

В [2] проводится анализ рисков возможных при реализации инновационных проектов.

В [199, 222] оценивается стоимость программного обеспечения на стадии разработки, в [132] построена модель для выявления дефектов в ПО, ненайденных в процессе тестирования.

1.3.5 Другие области применения. БСД также работают в других областях. В [172] описан способ их применения в анализе социальных сетей. В астрономии их применяют для классификации потоковых данных телескопа [206], в [131] — для диагностики работы спутника. В [226, 227] предложена модель, в которую

входят архитектурные элементы, решения архитектурного дизайна и связи между ними, что дает возможность архитекторам определять влияние изменений требований или дизайна.

1.4 ПОДХОДЫ К ОБУЧЕНИЮ БАЙЕСОВСКИХ СЕТЕЙ ДОВЕРИЯ

Параметры модели БСД можно разделить на параметры структуры и набор параметров (тензоры условной вероятности) вероятностного распределения. И соответственно этому обучение БСД сводится к двум задачам — локальное обучение параметров сети (структура сети известна) и глобальное обучение структуры.

1.4.1 Обучение структуры. Обучение структуры дает возможность реконструировать сеть на основе доступных данных, не зная заранее структуры и таблицы вероятностей байесовской сети. Создание точной структуры при обучении важно, так как отсутствие дуги приводит к неверной независимости, в то время как добавление дуги приводит к ложным зависимостям и увеличению числа параметров. Структура обучения байесовской сети объединяет априорные знания и выборку для определения структуры сети, которая наилучшим образом соответствует данным. Фундаментальным предположением при обучении структуры является то, что набор данных содержит независимо и идентично распределенные экземпляры, сгенерированные на основе базового распределения P, которое индуцируется байесовской сетью G [145].

В последние годы было разработано множество методов обучения структуры БСД. Алгоритмы обучения структуры БСД разделяют на три группы:

- Основанные на ограничениях (Constraint-based): эти алгоритмы изучают структуру сети, анализируя вероятностные отношения, обусловленные марковским свойством байесовских сетей и тестами на условную независимость, и затем строят граф, который удовлетворяет соответствующим утверждениям d-разделимости. Полученные модели часто интерпретируются как казуальные модели, даже если они обучены на данных наблюдений [195, 196].

- Метрические алгоритмы (score-based): эти алгоритмы дают оценку каждому кандидату байесовской сети и максимизируют ее с помощью некоторого эвристического алгоритма поиска. Алгоритмы жадного поиска (такие как восхождение к вершине или поиск по табу) являются наиболее распространенным выбором, но может быть использован почти любой вид поиска.
- Гибридные алгоритмы, то есть смесь первых и вторых [81].

1.4.1.1 Алгоритмы обучения структуры БСД, использующие ограничения. Алгоритмы обучения структуры БСД, использующие ограничения, основаны на алгоритме индуктивной причинности (IC), предложенный в [242], который обеспечивает теоретическую основу для изучения причинно-следственной структуры моделей. Его можно разделить на три этапа [214]:

1. Сначала обучается каркас сети (неориентированный граф, лежащий в основе структуры сети). Поскольку исчерпывающий поиск в вычислительном отношении невозможен практически во всех случаях, исключение составляют наиболее простые наборы данных, все алгоритмы обучения используют некоторую оптимизацию, такую как ограничение поиска «марковским одеялом» каждого узла (родители, дети и все узлы, которые совместно используют дочерний узел с этим конкретным узлом).
2. Определяются направления дуг, являющихся частью v-структуры (триплет узлов, образующих сходящуюся связь $X_j \rightarrow X_i \leftarrow X_k$).
3. Определяются направления остальных дуг исходя из ограничения ацикличности.

Алгоритмы обучения структуры БСД на основе ограничений имеют два недостатка. Одним из недостатков является экспоненциальное время выполнения относительно количества переменных. «Такие тесты могут быть ненадежны, если только их объем данных не огромен» [116]. Т. Верма и Дж. Перл также отметили, что множество утверждений об условной независимости будет расти экспоненциально по мере роста числа переменных [243]. Другим недостатком является их низкая устойчивость [178]. Небольшие изменения ввода могут сильно

изменить структуру байесовских сетей и создать ошибки в тестах на условные независимости.

Более эффективным алгоритмом является алгоритм PC [221], чья эффективность исходит от упорядочивания тестов на условную независимость от малых до больших. [220, 115].

Методы, основанные на ограничениях, такие как GS (Grow-Shrink) и TPDA (Three-Phase Dependency Analysis), часто начинаются с полного графа, а затем выполняются тесты на условную независимость (CI) для удаления как можно больше нежелательных ребер [176].

Наиболее недавние алгоритмы GS [178] и Inter-IAMB [251]. В их основе то же самое, но первые два шага используют более быстрые эвристики.

1.4.1.2 Алгоритмы обучения структуры БСД на основе метрик и поиска. В этом разделе будет описан один из самых популярных методов обучения байесовских сетевых структур — на основе наблюдаемых данных. Это алгоритмы эвристической оптимизации, которые ранжирует сетевые структуры по метрикам, таким как K2, BDe, AIC и BIC, чтобы оценить каждую подходящую структуру сети и попытаться найти оптимальную структуру, которая лучше всего подходит для данных выборки. Как следует из названия, Алгоритмы обучения структуры БСД на основе метрик и поиска, состоят из двух частей, одна из которых определяет метрику, которая позволяет оценивать различные сети. Вторая часть — это стратегия поиска, процедура оптимизации, которая позволяет определить наиболее подходящую структуру в пространстве поиска возможных сетевых структур.

Методы метрики и поиска предлагают возможные структуры и оценивают их, используя функцию оценки, которая измеряет, насколько хорошо байесовская сеть описывает набор данных D. Функция оценки должна измерить, насколько хорошо структура соответствует данным. Также желательно, чтобы она отбрасывала сложные структуры, потому что более простая модель делает вывод более понятным.

Метрические функции обычно подразделяются на две основные категории: функции байесовской оценки и информационно-теоретические скоринговые функции.

Байесовская оценка является одной из нескольких метрических функций, удовлетворяющих вышеупомянутым критериям. Предполагая структуру G, ее оценка равна $\mathrm{Score}(G, D) = P(G|D)$, другими словами, апостериорная вероятность *G* с учетом набора данных [178, 107]. Вычисление вышеупомянутого может быть приведено в более удобную форму с помощью правила Байеса: $P(G|D) = \frac{P(D|G)P(G)}{P(D)}$, где знаменатель является нормализующим фактором, который не зависит от оцениваемой структуры. Целью подхода, основанного на метрике, является максимизация оценки, которая присваивается каждому кандидату байесовской сети. Чтобы ее максимизировать, нужно только максимизировать числитель, поскольку знаменатель не зависит от G [178]. Тогда байесовская оценка выражается следующим образом: $\mathrm{Score}(G, D) = logP(D|G) + logP(G)$.

Дж. Купер и Э. Герсковиц [116] предложили байесовский метод, названный BLN (Bayesian learning of belief networks). Учитывая, что выполняется набор из четырех предположений, а именно: (i) переменные базы данных являются дискретными, (ii) события происходят независимо, учитывая модель сети доверия, (iii) данные не содержат пропусков, и, наконец, (iv) перед наблюдением базы данных, неважно размещение численных вероятностей на структуре сети доверия, — получается следующий результат [116, 218]:

Теорема. Пусть Z — это набор из n дискретных переменных, где переменная x_i в Z может иметь r_i значений $(v_{i1},...,v_{ir_i})$. Пусть D — это база данных из m случаев, содержащих значение для каждой переменной в Z. Пусть B_s — структура сети доверия, содержащая переменные из Z. Каждая переменная x_i в B_s имеет набор родителей $pa(i)$. Пусть w_{ij} обозначает *j*-ую уникальную

реализацию $pa(i)$ относительно D. Предположим, что есть q_i таких уникальных реализаций $pa(i)$. Определим N_{ijk} как число случаев в D, в которых переменная x_i имеет значение v_{ik}, а $pa(i)$ реализована как w_{ij}. Пусть $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Тогда при выполнении условий i–iv верно следующее: $P(B_S, D) = P(B_S)\prod_{i=1}^{n} g(i, pa(i))$, где $g(i, pa(i)) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!$.

Применяя данную теорему можно найти наиболее вероятную структуру сети при известных данных. Но при экспоненциальном росте количества структур за счет увеличения числа переменных найти наиболее вероятную структуру байесовской сети доверия становится невозможным.

Дж. Купер и Э. Герсковиц [147] предложили жадный алгоритм K2 [210], который максимизирует $P(B_s, D)$, определяя набор родителей для каждой переменной, который максимизирует $g(i, pa(i))$. Кроме условий i–iv, в этом случае предполагается также, что переменные упорядочены, а все структуры обладают одинаковой вероятностью. Алгоритм K2 описан ниже (листинг 1.1):

Где N_{ijk} вычисляются относительно π_i, являющимися родителями x_i и относящимися к базе данных D. Метрика K2 является одним из самых ранних алгоритмов для обучения структуры БСД и является частным случаем Байесовской Дирихле (BD) оценки [146].

Еще одна метрика, которую можно использовать, это метрика байесовского Дирихле (BDe) likelihood-equivalence Bayesian Dirichlet, получаемая с учетом вероятности и эквивалентности, дополнительные предположения о вероятностной эквивалентности и возможности структуры [107], ее выражение идентично выражению BD.

```
     [Ввод: множество n упорядоченных узлов, верхняя граница u
(максимальное число родителей узла), база данных D из m случаев. ]
     [Вывод: Для каждого узла выводятся его родители.]
     for i := 1 to n do
      pa(i): = ∅;
      Pold := g(i, pa(i));
      OKToProceed := true
      while OKToProceed and | pa(i)|< u do
                          пусть z — это узел-предшественник xi (исключая
            pa(i)), максимизирующий функцию g(i, pa(i) ∪ [z]);
            Pnew := g(i, pa(i) ∪ [z]);
            if Pnew > Pold then
                  Pold := Pnew;
                  πi := pa(i) ∪ [z]
            else OKToProceed := false;
      end [while];
      write('Узел:', xi, 'Родители узла: '; pa(i))
     end [for];
```

Листинг 1.1 — Алгоритм поиска структуры сети K2

Информационно-теоретические метрики основаны на сжатии. В этом контексте оценка байесовской сети G связана со сжатием, которое может быть достигнуто на данных D с оптимальным кодом, индуцированным G [107]. MDL-функция (minimum description length, вывод структур минимальных по длине описания) оценки предпочитает простые байесовские сети, а не сложные, и это строго определено как: $\mathrm{MDL}\left(G \mid D\right) = \mathrm{LL}\left(G \mid D\right) - 1log\left(N\right)|G|$, где $|X|$ обозначает сложность сети X, N обозначает общее количество экземпляров в данных и LL обозначает логарифмическую оценку вероятности. Обобщение функции подсчета MDL определяется следующим образом:

$\varphi\left(G \mid D\right) = \mathrm{LL}\left(G \mid D\right) - f\left(N\right) \mid G \mid$, где $f\left(N\right)$ — неотрицательная штрафная функция. Если $f\left(N\right) = 1$, есть функция критерия Akaike Information Criterion (AIC), то есть $AIC(G|D) = LL(G|D) - |G|$. Если $f\left(N\right) = 1log\left(N\right)$, то имеется Bayesian Information Criterion (BIC), основанный на информационном

критерии Шварца, совпадающем с оценкой MDL. Если $f(N) = 0$, может быть получена оценка LL [107].

В общем случае функции оценки должны быть разложимы по структуре сети в целях эффективности. Свойство разложимости позволяет использовать эффективные алгоритмы обучения, основанные на методах локального поиска. Более того, когда алгоритм обучения ищет в пространстве классов эквивалентности сетевых структур скоринговые функции также должны быть эквивалентным относительно оценок [112], то есть эквивалентные сети должны иметь одинаковую оценку.

Обучение БСД является довольно сложной задачей. Купер показал, что вывод общей байесовской сети представляет собой NP-сложную проблему, а позже, П. Дагум и М. Луби показали, что даже нахождение приближенного решения является NP-трудным [107]. Помимо функции оценки, следующее рассмотрение в парадигме «метрика и поиск» заключается в том, как найти структуру с высокой оценкой в пространстве структур. Стандартная методология для решения этой задачи — эвристический поиск, основанный на оптимизации метрик скоринга, который проводится на некотором пространстве поиска. Пространство поиска включает в себя сетевые структуры и классы эквивалентности упорядоченных сетевых структур по сетевым переменным.

Основные алгоритмы поиска в пространстве включают жадные алгоритмы поиска, генетические алгоритмы и алгоритмы имитации отжига.

Алгоритм жадного поиска, такой как восхождение на холм (hill-climbing) [111, 140], итеративно улучшает сеть, выполняя локальные модификации: добавление, удаление или изменение направления ребра. Поиск начинается с пустого, полного или возможно случайного графа. Предположим, есть некоторая структура сети, которая является направленным графом без циклов (DAG). Можно определить его окрестность в DAG-пространстве как все сети, к которым можно обратиться, применив оператор. Операторы поиска включают добавление ребер, удаление ребер и изменение направления ребер. При этом не должно появляться

циклов. На каждой итерации алгоритм вычисляет дельта-оценку (изменение в оценке) для каждой модификации и применяет одну модификацию с наибольшей положительной дельта-оценкой [107].

Процесс поиска останавливается при достижении локального максимума, где нет изменений, приводящих к улучшению оценки. Такой подход к обучению эффективен в вычислительном отношении и, хотя он не гарантирует оптимального результата, многие предыдущие исследования показали, что он получает очень хорошие решения. Алгоритмы восхождения на холм особенно популярны из-за их хорошего компромисса между вычислительными требования и качеством обученных моделей. Один из недостатков алгоритма восхождения на холм состоит в том, что обученная структура может быть локальной, а не глобальной, максимумой. Чтобы облегчить эту проблему, восхождение на гору часто дополняется случайным перезапуском или поиском табу [103]. Используя случайные перезапуски, делается некоторое количество случайных шагов, после застревания в локальном оптимуме, все начинается снова. Используя Табу-лист, можно сохранить список *K* шагов, совершенных последними, и не пересматривать недавно просмотренные структуры.

Пример жадного алгоритма восхождения к вершине:

1. Выбирается структура сети G над V , чаще всего пустая.

2. Вычисляется оценка Score_G.

3. $\mathrm{Score}_{\max} := \mathrm{Score}_G$.

4. Повторяются следующие шаги, пока увеличивается максимальная оценка:

(а) Для добавления дуги, удаления или обращения, не приводящего к циклической сети:

(1) вычисляется оценка модифицированной сети G^*, $\mathrm{Score}^*_G = \mathrm{Score}(G^*)$

(2) если $\mathrm{Score}^*_G > \mathrm{Score}_G$, то $G^* := G$ и $Score^*_G := Score_G$.

(б) $Score_{\max} := Score_G$.

5. Возвращается G.

Генетические алгоритмы имитируют естественную эволюцию посредством последовательного отбора «наиболее подходящей» модели и гибридизации их характеристик [167]. При этом пространство поиска исследуется через кроссовер (объединяющего структуру двух сетей) и мутацию (генерирующую случайные изменения) стохастических операторов.

Смысл роевых алгоритмов сводится к следующему: задается ряд представителей определенного роя. Они обладают начальной позицией в пространстве решений. Далее представители движутся в пространстве решений. Они используют правила, построенные на соотнесении индивидуальных знаний представителей и коллективном знании роя. Это позволяет выбирать наиболее оптимальные решения [14]. В качестве примеров таких алгоритмов можно назвать оптимизацию роем частиц [126], муравьиные алгоритмы [124, 125], пчелиные алгоритмы [90, 153] и др.

Алгоритмы имитации отжига (annealing algorithms) [103] выполняют стохастический локальный поиск, внося изменения, максимизирующие оценку сети, но при этом разрешая изменения, ее уменьшающие (с вероятностью, обратно пропорциональной уменьшению оценки) [214].

1.4.1.3 Гибридные алгоритмы. Гибридные методы сочетают в себе особенности методы на основе ограничений и метрик. На первом этапе проверка на независимость используется для построения каркаса БСД, чтобы уменьшить пространство поиска. На втором этапе используются методы метрик и поиска для обнаружения направленного ациклического графа, оптимизирующего скоринговую функцию исходной байесовской сети [175]. М. Сингх и М. Вальторта [218] интегрировали тесты на условную независимость (CI) для генерации упорядочения узлов и байесовской оценки для воссоздания сетевых структур. Д. Дэш и М. Друздзель в [121] предложили искать пространство эквивалентных классов основных графов с использованием эвристических подходов, основанных на ограничениях, а затем оценивать с помощью байесовской метрики.

С. Ацид и Л.М. де Кампос [82] разработали метрику оценки на основе расхождений и эвристику стратегия поиска, которая подчеркивает баланс между сложностью модели и точностью.

И. Цамардинос предложил алгоритм Max-Min Hill-Climbing (MMHC), который объединил методы локального обучения и ограничения, строя ненаправленный граф, и затем выполняя жадный поиск с помощью Байесовской метрики, чтобы придать направление дугам [240].

В 2017-ом Хуэй Лю и Шуйгенг Чжоу разработали новый гибридный метод SAR (разделение и воссоединение) [176], основанный на декомпозиции неориентированного графа независимости на полный набор узлов на этапе разделения и обучении небольших направленных ацикличных графов для набора узлов в соответствии с каждым подграфом путем применения метода оценивания на этапе воссоединения.

1.4.2 Обучение параметров. Если состояния всех узлов сети дискретны, чтобы представить причинные отношения между узлом и его родителями, используются таблицы условной вероятности. Обучение параметров при известной структуре сети является задачей оценивания всех тензоров вероятности на основе наблюдаемых данных [154].

БСД дают решения для экспертных систем, основываясь на тензорах условной вероятности и структуре сети. Если тензоры условной вероятности неизвестны, единственный способ получить условные вероятности параметров из наблюдаемых данных. Алгоритмы обучения параметров БСД можно разделить на алгоритмы для полного набора данных и для неполного набора данных. Обсудим наиболее используемые алгоритмы из этих категорий.

Метод максимального правдоподобия (Maximum Likelihood Estimate, MLE) — наиболее частая стратегия для полного набора данных. [136]

Основная идея MLE: случайный тест для события (C) может вызвать несколько возможных результатов C^1, C^2, ..., C^n. Если результатом этого теста является C^k $(1 \le k \le n)$, оно будет максимальным правдоподобием для этого

события. Следовательно, оценочное значение $\hat{C}$ будет установлено как параметр θ , если оно может максимизировать значение функции правдоподобия.

Функция правдоподобия для байесовской сети с n узлами определяется следующим образом:

$$L(\Theta : D) = \prod_m P(x_{1m}, \ldots, x_{nm} \mid \Theta) = \prod_m \prod_i P(x_{im} \mid pa(i)_m, \Theta_i) =$$

$$= \prod_i \prod_m P(x_{im} \mid pa(i)_m, \Theta_i) = \prod_i L_i(\Theta_i : D).$$

Если вероятность $P(x_i \mid pa(i))$ удовлетворяет полиномиальному распределению, локальная функция правдоподобия может быть далее разложена:

$$L_i(\theta_i : D) = \prod_m P(x_{im} \mid pa(i)_m, \theta_i) = \prod_m \prod_{pa(i)^j} \prod_{x_i'} P(x_{im}^k \mid pa(i)_m^j, \theta_i) = \prod_{pa(i)^j} \prod_{x_i'} \theta_{x_i' pa(i)}^{N(x_i', pa(i)^j)}.$$

Учитывая, что набор данных полный, для каждого возможного значения $pa(i)^j$ узлов родителей $pa(i)$, распределение вероятности $P(x_i \mid pa(i)^j)$ — это независимое по

$$L_i(\theta_i : D) = \prod_m P(x_{im} \mid pa(i)_m, \theta_i) = \prod_m \prod_{pa(i)^j} \prod_{x_i'} P(x_{im}^k \mid pa(i)_m^j, \theta_i) = \prod_{pa(i)^j} \prod_{x_i'} \theta_{x_i' pa(i)}^{N(x_i', pa(i)^j)}.$$

линомиальное распределение, которое не относится ко всем остальным значениям $pa(i)^l (l \neq j)$, тогда MLE может получить оценку параметра θ как:

$$\Theta_{x_i^k} = \frac{N(x_i^k, pa(i)^j)}{N(pa(i)^j)}.$$

Исходя из этого выражения, можно легко получить таблицы условных вероятностей при известной топологии сети. На самом деле, метод MLE использует частоту вместо вероятности. Он сходится медленно из-за отсутствия каких-либо предварительных знаний в процессе оценки параметра.

Основная идея байесовского метода [136] для обучения параметров заключается в следующем: дано распределение с неизвестными параметрами и полный набор (C) данных, θ является случайной величиной с априорным распределением $p(\Theta)$; изменения параметра θ, а именно $p(\Theta \mid C)$, может быть

оценен в соответствии с предыдущими знаниями или предположением, что у $p(\theta)$ равномерное распределение. Поэтому $p(\theta \mid C)$ называется апостериорной вероятностью параметра θ. Цель этого метода состоит в том, чтобы вычислить эту апостериорную вероятность, что будет рассматриваться как основа обучения параметра.

Предположим, что априорное распределение $p(\theta)$ — это распределение Дирихле:

$$p(\theta) = Dir(\theta \mid \alpha_1, ..., \alpha_r) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k^{-1}},$$

где $\alpha = \sum_{k=1}^{r} \alpha_k$, $\alpha_k > 0$, $k = 1, ..., r$, $\alpha_1, ..., \alpha_r$ являются гиперпараметрами; Гамма-функция $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$ удовлетворяет $\begin{cases} \Gamma(x+1) = x\Gamma(x) \\ \Gamma(1) = 1 \end{cases}$. Отсюда:

$$p(D) = \int p(\theta) p(D \mid \theta) d\theta = \int \frac{\Gamma(\alpha)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k^{-1}} \times \prod_{k=1}^{r} \theta_k^{N_k} d\theta = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^{r} \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)}$$

Процесс обучения параметров в БСД можно рассмотреть в трех ситуациях:

1. Расчет вероятности узлов, у которых нет родителей.

$$p(\theta \mid D) = \frac{p(\theta) p(D \mid \theta)}{p(D)} = \frac{\Gamma(\alpha + n)}{\prod_i \Gamma(\alpha_i + n_i)} \prod_k \theta k^{\alpha_i - n_i} = Dir(\alpha_1 + n_1, ..., \alpha_N + n_N).$$

Тогда $p(\theta_i \mid D) = \frac{\alpha_i + n_i}{\alpha + N}$, где n_i — это число вхождений i-ых возможных значений x_i, а N — число вхождений всех возможных значений переменной x_i во множестве данных.

2. Расчет условной вероятности для узлов, у которых есть только один родитель.

Предполагается, что связь между узлами X и Y можно обозначить как $X \to Y$, и они являются дискретными переменными. Следовательно, $p(y \mid x_i, \theta) = Dir(\alpha_{i1} + n_{i1}, ..., \alpha_{ik} + n_{ik})$.

3. Расчет условной вероятности для узлов, которые имеют несколько родителей.

Во-первых, делается предположение о независимости параметров: параметры, которые могут иметь различное распределение, взаимно независимы. θ_{ijk} представляется как условная вероятность, когда $pa(i) = j$ и $x_i = k$. Значение θ_{ijk} может быть вычислено следующим образом:

$$\theta_{ijk} = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \bullet \left(\alpha_{ij} = \sum \alpha_{ijk}, n_{ij} = \sum n_{ijk} \right).$$

Основная идея метода EM (Expectation Maximization) [168]: когда наблюдаемые данные неполные, можно использовать алгоритмы вывода байесовской сети для оценивания отсутствующего значения набора данных, чтобы набор данных стал полным. EM-алгоритм включает в себя два этапа: (1) инициализация: θ_s присваивается случайное значение; (2) расчет исключения. Рассчитывается статистический коэффициент ожидания каждого события « $X_i = k, pa(i) = j$ » при условии θ_s; Для дискретных переменных $E_{p(X|D,\theta_s,S^h)}(N_{ijk}) = \sum_{i=1}^{N} p(x_i = k, pa(i) = j \mid y_i, \theta_s, S)$, где N_{ijk} является достаточным статистическим коэффициентом события « $X_i = k, pa(i) = j$ »; (3) максимизация исключения. Коэффициент достаточности исключения был применен для преобразования неполных данных D в полные выборки данных; (4) если точность достигнута, остановка процесса; в противном случае возврат к (2). Следовательно, можно вычислить условную вероятность через выражение:

$$\theta_{ijk} = \frac{E_{p(x|D,\theta_s,S^h)}(N_{ijk})}{\sum_{k=1}^{r_i} E_{p(x|D,\theta_s,S^h)}(N_{ijk})}.$$

Подход EM использует алгоритм байесовского сетевого вывода для вычисления исключения переменных, которые не наблюдаются. Выбор алгоритма вывода является ключом для подхода EM. Алгоритм Junction Tree [152] широко используется в EM-подходе для обучения параметров.

Стратегия отбора Гиббса, относится к довольно известному методу Монте-Карло [192, 229]. Основная идея такова: ожидание $f(X)$ совместного распределения вероятностей $P(X)$ на множестве переменных X оценивается с помощью выборки Гиббса. Этот метод включает в себя следующие шаги: (1) инициализация. Произвольная инициализация осуществляется на неполном наборе данных D, чтобы получить полный набор данных D_C. (2) Выбирается ненаблюдаемая переменная X_{im} в наборе данных D (*m*-й образец экземпляра *i*-й переменной) и случайным образом определяется следующим распределением вероятностей:

$$p\left(X'_{im} \mid D_C \setminus X_{im}, S\right) = \frac{p\left(X'_{im} \mid D_C \setminus X_{im}, S\right)}{\sum_{X^n_{im}} p\left(X^n_{im} \mid D_C \setminus X_{im}, S\right)}.$$

Где $D_C \setminus X_{im}$ обозначает ситуацию перехода наблюдения от X_{im}; (3) на шаге 2 каждая ненаблюдаемая переменная в *D* была определена, таким образом получен новый рандомизированный полный набор данных D'_C. (4) вычисляется $p(\theta_s \mid D'_C, S)$; (5) шаги 1–4 повторяются столько раз, сколько нужно, чтобы получить среднее значение всех $p(\theta_s \mid D'_C, S)$.

Методы Монте-Карло очень гибкие, когда другие методы не применимы. Чем больше выборка, тем точнее результаты; однако вычислительная сложность экспоненциальна числу экземпляров выборки. Выборка Гиббса является наиболее типичным методом Монте-Карло.

1.5 ИНСТРУМЕНТЫ ДЛЯ РАБОТЫ С БАЙЕСОВСКИМИ СЕТЯМИ ДОВЕРИЯ

Кевин Мерфи в [184] представляет довольно обширный список (70 проектов) программного обеспечения для работы с БСД, но некоторые пункты представляют собой проекты, которые больше не поддерживаются (XBAIES [250], BayesBuilder [91], Elvira [128]).

Одним из важнейших свойств БСД является их высокая интерпретируемость за счет отображения причинно-следственных связей между элементами модели, поэтому в инструментах, работающих с БСД, важно наличие средств для визуализации моделей. Еще один важный аспект для таких инструментов — возможность пошагово создавать модели, не вникая при этом в технические детали. Такую возможность предоставляют инструменты с графическим интерфейсом, они позволяют создавать структуры сетей легко и понятно для пользователя. Ниже приведены примеры соответствующих инструментов.

AgenaRisk [83]. Компания Agena Ltd была создана как консалтинговая компания в 1998 году на основе исследований, посвященных использованию крупномасштабных БСД в различных областях, что в итоге привело к созданию программного продукта AgenaRisk для анализа рисков и принятия решений. AgenaRisk предоставляет платные продукты с 14-дневной пробной версией, также для ученых, студентов и исследователей предоставляется скидка.

BayesiaLab [93]. Лидирующее настольное программное обеспечение для работы с БСД. Мощное настольное приложение со сложным графическим интерфейсом пользователя, является одним из лидирующих инструментов в этой области, благодаря BayesiaLab работать с БСД могут работать исследователи во многих областях, не связанных напрямую с компьютерными науками. Это платное ПО, с 30-дневной, также функционально- ограниченной пробной версией.

Bayes Server [95]. Компания Bayes Server начинала с разработки решений в области искусственного интеллекта для General Electric (GE) и ВВС США (USAF), и сейчас имеет более чем 15-летний опыт поставки готового и индивидуального

программного обеспечения в области искусственного интеллекта для самых передовых компаний в мире. Пользовательский интерфейс пока разработан только для системы Windows, однако с API можно работать с различных платформ. Это также платный продукт.

BayesFusion [92] предоставляет программное обеспечение для моделирования искусственного интеллекта и машинного обучения на основе БСД. Работать с ПО этой компании можно на настольных компьютерах, мобильных устройствах и в облаке.

Основной продукт компании — это GeNIe Modeler, инструмент для моделирования искусственного интеллекта и машинного обучения с использованием байесовских сетей и других типов графических вероятностных моделей.

GeNIe Modeler является графическим интерфейсом программной библиотеки SMILE Engine, которая позволяет работать с БСД из клиентских приложений, которые могут быть написаны на различных языках программирования (например, C++, Python, Java, .NET, R, Matlab).

Модели, созданные с помощью GeNIe и/или SMILE, можно легко распространять и использовать на мобильных устройствах с помощью BayesMobile или через веб-браузер с помощью BayesBox. Большим достоинством этих продуктов является бесплатное использование для обучения и исследовательских целей.

Кроме обычной функциональности, с помощью GeNIe можно предварительно обрабатывать данные (например, разбивать значения на дискретные интервалы), строить структуру сети на основе данных, рассчитывать различные метрики сети.

Hugin Expert [149]. Компания HUGIN EXPERT A/S и была основана в 1989 году ведущими мировыми исследователями в области графических моделей, основанных на технологии БСД в университете Ольборга, в Дании. Есть версия с графическим интерфейсом, а также возможность работы с API, программа является платной.

Netica [187] — это мощная, простая в использовании, программа для работы с БСД и диаграммами влияния. Она имеет интуитивно понятный пользовательский интерфейс для построения сетей, а связи между переменными могут быть введены как отдельные вероятности, в виде уравнений или получены из файлов данных (представленными в табличном виде и иметь недостающие данные). Приложение Netica и Netica API, находятся в постоянной разработке с 1992 года, с 1995 стали доступны для коммерческого использования. Их разрабатывает компания Norsys, зарегистрированная с 1996 года. Это платное ПО, однако есть бесплатная версия, которая является полнофункциональной, но имеет ограниченный размер модели. В Netica также есть дополнительные интересные функции, например, генерация веб-сайтов на основе БСД, высококачественной графики, сетевых отчетов по различным свойствам сети, средства для легкой дискретизации непрерывных переменных, поддержка отсоединенных связей и др.

BUGS (Bayesian inference Using Gibbs Sampling) [104]. Цель проекта — создание гибкого программного обеспечения для байесовского анализа сложных статистических моделей с использованием методов Марковской цепи Монте-Карло. Он начался в 1989 году в отделе биостатистики MRC (Кембридж), процессе сначала была создана программа BUGS, а затем программное обеспечение с графическим интерфейсом WinBUGS, разработанное совместно с Медицинской школой Имперского колледжа Сент-Мэри (Лондон). В дальнейшем развивалось ПО OpenBUGS, эквивалент WinBUGS с открытым исходным кодом.

Кроме указанных выше программ существует множество других средств, обеспечивающих работу с БСД. Часть из них является библиотеками, пакетами и фреймворками для других программ и языков программирования: пакеты для R (bnlearn [98], gRain [143], deal [122], catnet [109],), библиотеки для Java (Dimple [123], Weka [247]), для C++ (libDAI [174], OpenGM2 [193]), для C# (Infer.NET [150]), для Python (PyMC [197]), для Scala (Bayes-Scala [94], Factory [129]), для Matlab (BNL [97], BNT [99]) etc.

Стоит также отметить Stan [223], это язык вероятностного программирования, реализующий полный байесовский статистический вывод,

приближенный байесовский вывод с вариационным выводом и оценку максимального правдоподобия с оптимизацией. Математическая библиотека Stan предоставляет дифференцируемые функции вероятности и линейную алгебру, а дополнительные пакеты для R и Python обеспечивают линейное моделирование на основе экспрессии, визуализацию апостериорных данных и перекрестную валидацию по оставленным значениям. Stan — это свободное программное обеспечение с открытым исходным кодом (новое ядро BSD, некоторые интерфейсы GPLv3).

В предлагаемом исследовании одним из наиболее значимых инструментов, обеспечивающих работу с БСД является пакет bnlearn [98] для R, это пакет для изучения графической структуры БСД, оценивания их параметров и выполнения некоторых других действий. Он был впервые выпущен в 2007 году, он находится в постоянном развитии уже более 10 лет (и все еще продолжает развиваться).

1.6 ПОСТАНОВКА ЦЕЛИ И ЗАДАЧ ИССЛЕДОВАНИЯ

При сборе информации о поведении человека возникают множество источников неопределенности, часть из которых относятся к предметной области (не рассматриваются в диссертационном исследовании), например, факторы, оказывающие влияние на конкретный тип поведения, а часть являются общими для задачи оценивания характеристик процесса по ограниченным данным самоотчетов.

Задача диссертационного исследования заключается в развитии подхода, предложенного ранее в рамках исследований лаборатории теоретических и междисциплинарных проблем информатики [30–32], ранее в этом подходе не рассматривались следующие виды неопределенности:

1) Несогласованность ответов респондентов, то есть рассматриваются ситуации, когда респондент дает ответы, противоречащие друг другу. Например, если на вопросы о физических нагрузках: «Когда Вы занимались физическими упражнениями в последний раз?» и «Каков Ваш минимальный интервал между занятиями физическими упражнениями?» респондент ответит, что в последний раз

он занимался спортом вчера, перед этим позавчера, и укажет при этом минимальный интервал в 2 дня, то эта информация будет не согласована.

2) Некорректность ответов респондентов. Под некорректностью ответов респондентов, понимается их расхождение с реальным положением дел, причем такое искажение информации может быть как намеренным (например, если собирается информация о социально одобряемых видах поведения) или ненамеренным (если эпизоды процесса произошли какое-то время назад, и точную дату быстро припомнить сложно).

3) Некорректность задания момента окончания исследования (момент сбора информации). В ряде ситуаций момент сбора информации об эпизодах поведения зависит от самого процесса поведения. Кроме того, при фиксации момента сбора информации в базе данных могут быть допущены ошибки.

В рамках решаемой задачи цель исследования сформулирована следующим образом: повысить качество косвенной оценки интенсивности пуассоновского процесса как математической модели эпизодического поведения индивида за счет разработки методов и алгоритмов обработки некорректности и неопределенности данных, предоставляемых респондентами.

Для выполнения поставленной цели необходимо усовершенствовать существующие модели оценки интенсивности пуассоновского процесса или разработать новые с учетом рассматриваемых видов неопределенности.

Для удобной работы с описанными моделями и методами, а также для возможности их использования в различных областях исследования в дальнейшем, необходимо автоматизировать работу, в частности реализовать компоненты прототипа комплекса программ для работы с указанными моделями, их анализа, апробации, вычислительных экспериментов и решения практических задач.

ВЫВОДЫ ПО ГЛАВЕ 1

В первом разделе обоснована актуальность исследования. Во втором разделе обосновано использование БСД для решения поставленной задачи, а также сделан обзор сфер, в которых они применяются. Третий раздел посвящен подходам к

обучению БСД. В четвертом разделе сделан обзор инструментов для работы с БСД. В пятом разделе поставлена цель и задачи исследования.

Во многих областях исследования более точная оценка интенсивности пуассоновского процесса за счет обработки возникающей при этом неопределенности данных может повысить качество результатов, таким образом результаты, полученные в ходе решения поставленной задачи, могут найти широкое практическое применение.

БСД способны эффективно комбинировать несколько видов информации (к примеру, статистическую и экспертную). Такая возможность делает БСД инструментом понятным, удобным, производительным. Могут возникнуть трудности с построением адекватной модели некоторого процесса, в этом случае можно воспользоваться методами обучения структуры, при этом благодаря свойствам этого инструмента статистические данные можно совмещать с экспертной информацией. В том же случае, когда структура БСД определена при достаточном количестве данных вместо экспертных предположений или расчетов тензоров условной вероятности можно использовать обучение параметров сети.

БСД является подходящим инструментарием для оценивания интенсивности пуассоновских процессов, что также было показано в работах А.Л. Тулупьева, А.Е. Пащенко, А.В. Суворовой, Т.В. Тулупьевой, А.В. Сироткина, С.И. Николенко [21, 30–32].

ГЛАВА 2. ОСНОВНЫЕ ПОНЯТИЯ И ИСПОЛЬЗУЕМЫЕ МЕТОДЫ

Вторая глава посвящена описанию теоретических результатов, послуживших предпосылками для данного диссертационного исследования и создающих основу для решения поставленных задач. Описаны основы теории БСД. Представлен разработанный ранее подход к оцениванию интенсивности пуассоновского процесса как модели эпизодического поведения индивида по сверхкороткому набору наблюдений с использованием БСД. Также описываются используемые метрики качества моделей и структура исследования. Данная глава не включает результаты данного диссертационного исследования, а предназначена для введения единой системы обозначений и описания предложенного ранее подхода к моделированию пуассоновских процессов, в развитии которого заключается суть диссертационного исследования.

2.1 ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ БАЙЕСОВСКИХ СЕТЕЙ ДОВЕРИЯ

Теория БСД в конечном счете берет свое начало из работы Томаса Байеса, впервые сформулировавшего эвристику: «апостериорная вероятность события прямо пропорциональна его правдоподобию» [69].

Байесовская сеть доверия — это вероятностная графическая модель, которая представляет собой ациклический направленный граф (в таком графе не может быть направленных циклов, но могут быть ненаправленные), его вершинами являются случайные элементы, входящие в модель, а ребра обозначают причинно-следственные связи между элементами, помимо графа должно быть заданы тензоры условной вероятности, определяющие переходы между вершинами [63–69, 185, 195]. Существуют двоичные, многозначные, непрерывные случайные элементы, заданные посредством распределений вероятности. В данном диссертационном исследовании используются БСД с многозначными случайными элементами.

Правомерно и такое определение: «БСД — это пара ⟨G,P⟩, где G — ациклический направленный граф (то есть граф, в котором нет направленных циклов, ненаправленные циклы допускаются), P — совместное распределение вероятностей всех случайных элементов, приписанных узлам графа (множество тензоров условных вероятностей), причем для любого узла соответствующий ему случайный элемент условно независим от всех своих непотомков, которые также не являются его родителями, при заданном означивании своих родителей» [69].

В графе G два случайных элемента являются условно независимыми по отношению к элементам, получившим означивание, если в любом (в том числе ненаправленном) пути между узлами (a и b), соответствующими этим двум элементам, есть узел c такой, что верно одно из условий (рисунок 2.1):

1) Оба ребра пути выходят из c (расходящаяся связь);

2) Одно ребро пути выходит из c, а другое входит (последовательная связь);

3) Узел c и его потомки не получили означивание, и оба ребра пути входят в c (сходящаяся связь).

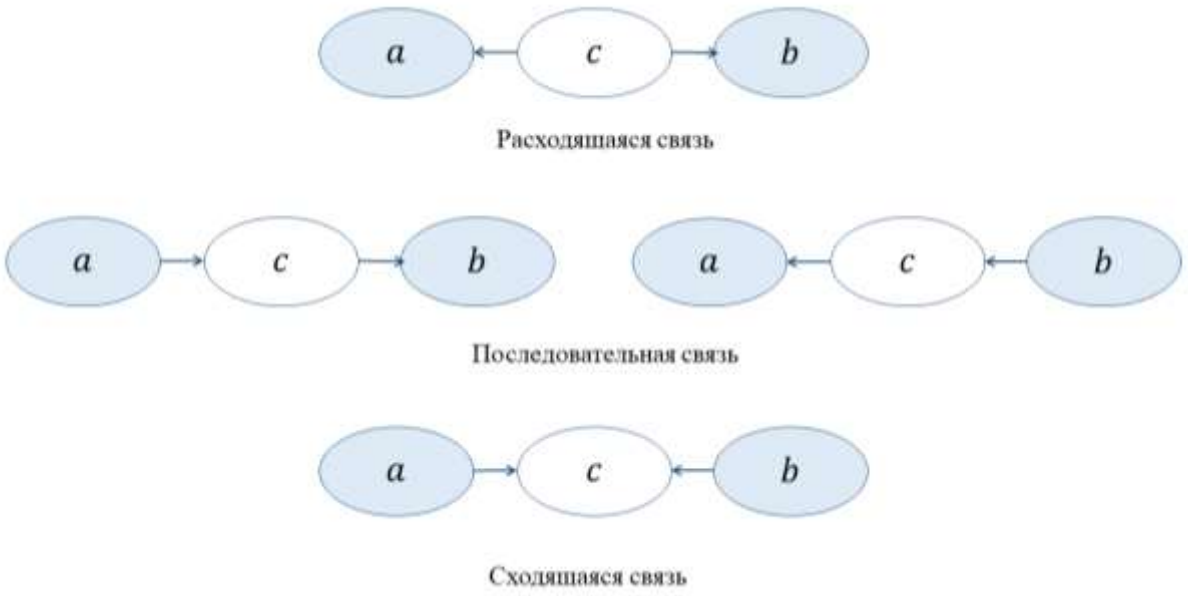

Рисунок 2.1 — Варианты связей между вершинами БСД

Если в каком-либо пути между узлами a и b есть такой c, то говорят, что c блокирует этот путь при заданной совокупности элементов, получивших

означивание. Если при этом блокированы все пути между a и b, говорят, что заданная совокупность элементов, получивших означивание, d-разделяет узлы a и b.

На основе понятия d-разделимость можно дать еще одно определение БСД: «ациклические направленные графы с тензорами условных вероятностей в узлах, причем случайные элементы, соответствующие d-разделенным узлам, являются условно независимыми при заданных означиваниях случайных элементов, соответствующим узлам из d-разделяющего множества».

«Вероятностной семантикой БСД при любом определении является то единственное распределение вероятностей, которое эта сеть задает» [64, 65, 68, 69].

Одно из основополагающих свойств БСД, это свойство декомпозиции, которое формулируется следующим образом: $P(X_1,...,X_m)=\prod_{i=1...m} P(X_i \mid \mathrm{pa}(X_i))$, где $P(X_1,...,X_m)$ — общее распределение вероятностей всех случайных элементов, $P(X_i \mid \mathrm{pa}(X_i))$ — распределение вероятностей случайного элемента X_i при означивании случайных элементов $\mathrm{pa}(X_i)$ (родители вершины X_i).

Означивание случайного элемента (то есть этот случайный элемент принял определенное значение из возможных) называют свидетельством.

При поступлении свидетельств в БСД можно определить апостериорные распределения случайных элементов, входящих в модель (провести апостериорный вероятностный вывод).

Примеры различных алгоритмов вывода содержатся в [68, 69, 185, 195, 196], эти алгоритмы реализованы в программном обеспечении, работающем с БСД (см. раздел 1.5).

2.2 СУЩЕСТВУЮЩИЕ МОДЕЛИ ОЦЕНИВАНИЯ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА НА ОСНОВЕ БАЙЕСОВСКИХ СЕТЕЙ ДОВЕРИЯ

Основная задача модели [30, 32] — получение оценки интенсивности поведения, в частности рискованного, по данным устного опроса или анкетирования респондентов [71] о последних эпизодах поведения и рекордных интервалах между последовательными эпизодами за определенный промежуток времени.

Пуассоновский процесс является естественной моделью для рассматриваемого эпизодического поведения, отвечающего следующим предположениям: эпизоды поведения происходят в непрерывном времени, за конечный промежуток может произойти только конечное число эпизодов, эпизоды не могут происходить одновременно, время реализации эпизода для каждого индивида не зависит от времени предыдущих эпизодов, и интенсивность поведения индивида остается постоянной во времени. Гамма-пуассоновская модель поведения возникает при учете неоднородности популяции в плане эпизодического поведения: интенсивность поведения варьируется между индивидами и моделируется гамма–распределенной случайной величиной.

На рисунке 2.2 представлена обобщенная модель байесовской сети доверия $M = \left(G(V,L), \mathbf{P}\right)$ [67, 195]. Ее структура выражена графом $G(V,L)$, где $V = \left\{t_{01}, t_{12}, t_{23}, t_{\min}, t_{\max}, \lambda, n\right\}$ — множество вершин, $L = \left\{(u,v) : u,v \in V\right\}$ — множество направленных связей между вершинами [56].

λ — это случайная величина, описывающая интенсивность (частоту) поведения. В качестве ее вероятностного распределения используется гамма распределение вероятности [8]. Еще один вариант — расчет тензоров условной вероятности по собранным статистическим данным. t_{ij} — случайная величина, которая соответствует длине интервала между *i*-ым и *j*-ым эпизодами поведения (для простоты будем полагать, что i может принимать значения 1 и 2, а $j = i+1$).

Она имеет экспоненциальное распределение, исходя из допущения, что поведение может рассматриваться как пуассоновский процесс. Интервал между временем получения информации от респондента и последним эпизодом не является интервалом между последовательными эпизодами поведения, в регрессионном анализе для моделирования этой особенности используется используется бета-простое распределение [8]. Участие в модели рекордных ($t_{\min}$ и $t_{\max}$) интервалов позволяет получить дополнительную информацию. Отметим, что минимальный и максимальный интервалы поведения за определенный промежуток времени T. n — это скрытая переменная, которая соответствует числу эпизодов поведения за отрезок времени T. Все непрерывные случайные величины, входящие в модель, полагаются дискретизированными.

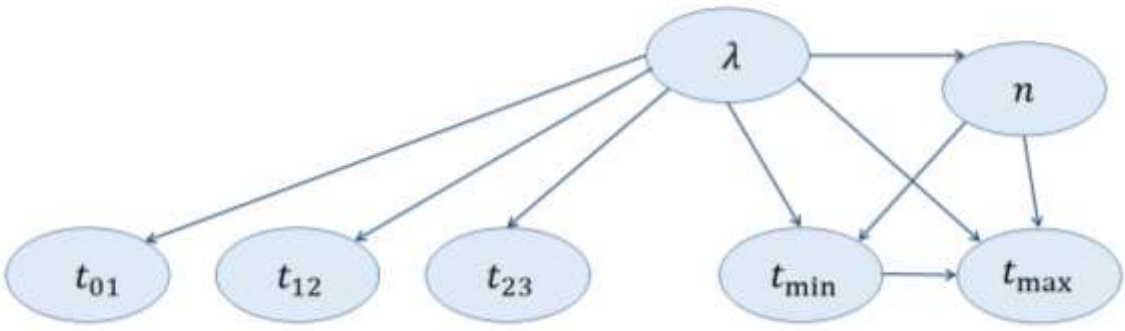

Рисунок 2.2 — Модель оценивания интенсивности пуассоновского процесса [31]

Тензоры условной вероятности $\mathbf{P} = \left\{ P\left(t_{j,j+1} \middle| \lambda\right), P\left(t_{01} \middle| \lambda\right), P\left(t_{\min} \middle| n, \lambda\right), P\left(t_{\max} \middle| n, \lambda, t_{\min}\right), P\left(n \middle| \lambda\right), P\left(\lambda\right) \right\}$, задающие переходы между узлами, определены следующим образом [32]:

$$p\left(t_{j,j+1}^{(l_j)} \middle| \lambda^{(i)}\right) = e^{-a\lambda^{(i)}} - e^{-b\lambda^{(i)}},\ j = 0,1,2,\ t_{j,j+1}^{(l_j)} = [a;b)\ ;$$

$$p\left(t_{\min}^{(l_3)} \middle| n, \lambda^{(i)}\right) = e^{-an\lambda^{(i)}} - e^{-bn\lambda^{(i)}},\ t_{\min}^{(l_3)} = [a;b);$$

$$p\left(n \middle| \lambda^{(i)}\right) = \frac{\left(\lambda^{(i)} T\right)^n}{n!} e^{-\lambda^{(i)} T};$$

$$p\left(t_{\max}^{(l_4)}\middle|n,\lambda^{(i)},t_{\min}^{(l_3)}\right)=e^{(n-1)\lambda^{(i)}t_{\min}^{(l_3)}}\left(\left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}}-e^{-\lambda^{(i)}b}\right)^{n-1}-\left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}}-e^{-\lambda^{(i)}a}\right)^{n-1}\right),$$

$t_{\max}^{(l_4)}=[a;b)$;

$l_s=1,...,k_s$, где k_s — число дизъюнктных интервалов при дискретизации случайных величин; $s=0,...,4$; $j=1,...,2$; $i=1,...,m$, где m — число дизъюнктных интервалов при дискретизации λ.

После того как тензоры условных вероятностей для всех узлов определены, численные расчеты апостериорного распределения частоты поведения при условии входных статистических данных могут быть выполнены с помощью программного обеспечения (см. раздел 1.5).

Результаты апробации модели оценивания интенсивности пуассоновского процесса представлены в [30].

В работе [31] также была представлена модель со структурой, построенной по синтетическим данным (рис. 2.3). При этом был использован алгоритм Hill-Climbing (HC) [213], за меру качества брался BIC (Bayesian Information Criterion).

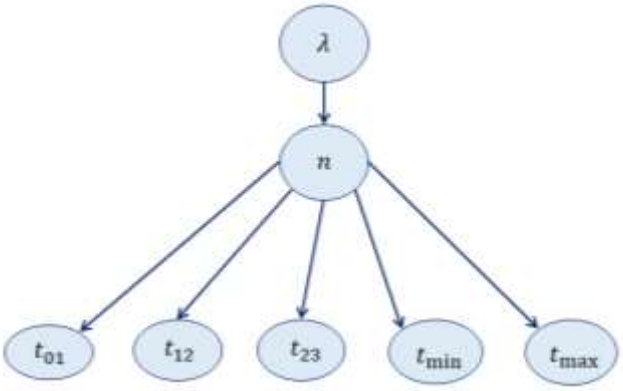

Рисунок 2.3 — Структура БСД, обученная на синтетических данных [31]

Такой структуре, обученной на данных, можно дать такое объяснение: λ определяется по количеству эпизодов поведения за исследуемый период n, которая в свою очередь определяется по исходным данным. Стоит отметить, что значение переменной n как и λ в большинстве случаев поведения невозможно получить прямыми методами, таким образом переменные n и λ являются скрытыми.

При оценивании интенсивности пуассоновского процесса как модели эпизодического поведения на основе моделей БСД по данным респондентов в работе рассматривается неопределенность следующих видов:

1. Несогласованность ответов респондентов.

2. Некорректность ответов респондентов.

3. Некорректность задания момента окончания исследования.

Под некорректностью ответов респондентов, понимается их расхождение с реальным положением дел, а под несогласованностью информации понимаются ситуации, когда респондент дает ответы, противоречащие друг другу. Для обработки этих видов неопределенности в диссертационной работе предлагаются решения.

2.3 ПОКАЗАТЕЛИ КАЧЕСТВА МОДЕЛЕЙ КЛАССИФИКАЦИИ НА ОСНОВЕ БАЙЕСОВСКИХ СЕТЕЙ ДОВЕРИЯ

В данном диссертационном исследовании используются БСД с многозначными случайными элементами в узлах сети. Таким образом задача оценивания интенсивности пуассоновского процесса сводится к задаче классификации, поэтому для определения качества моделей в диссертационном исследовании используется матрица ошибок и выводимые из нее метрики.

Матрица ошибок представляет собой таблицу, в которой в строках указаны действительные значения некоторого фактора, а в столбцах — полученные с помощью некоторого классификатора оценки. Таблица 2.1 представляет собой пример такой таблицы, где m — это количество классов некоторого фактора, а y_{ij} представляет собой количество экземпляров i-ого класса, отнесенных классификатором к классу j. Такая таблица дает общее представление об эффективности модели классификации.

При этом для каждого класса i можно вычислить: количество истинно-положительных наблюдения (True Positive, TP) — это случаи, правильного определения соответствующих классов, $\mathrm{TP}_i = y_{ii}$; количество истинно-

отрицательных наблюдений (True Negative, TN) — это наблюдения, правильно не включенные в классы, $\mathrm{TN}_i = \sum_{j=1, j\neq i}^{m} y_{ji}$; количество ложно-положительных наблюдений (False Positive, FP) — это случаи, неправильно включенные в классы, $\mathrm{FP}_i = \sum_{j=1, j\neq i}^{m} y_{ij}$; количество ложно-отрицательных наблюдений (False Negative, FN) — это наблюдения, неправильно не включенные в классы, $\mathrm{FN}_i = \sum_{j,k=1; j,k\neq i}^{m} y_{jk}$. Таким образом TP и TN — это случаи правильной работы классификатора, а FN и FP — это случаи неверной работы классификатора [17].

Таблица 2.1 — Матрица ошибок

	Оценки классификатора			
Действительные значения	y_{11}	y_{12}	…	y_{1m}
	y_{21}	y_{22}	…	y_{2m}
	…	…	…	…
	y_{m1}	y_{m2}	…	y_{mm}

Точность (accuracy) — отношение правильно сделанных оценок к общему числу значений. Ее можно вычислить, поделив сумму значений, находящихся на главной диагонали (из левого верхнего угла в правый нижний), на сумму всех значений в матрице ошибок.

$\mathrm{Ac} = \frac{\sum_{i=1}^{m} \mathrm{TP}_i}{N}$, где N — общее число выборки ($N = \sum_{i,j=1}^{m} y_{ij}$).

Несмотря на большую важность этой метрики, возможна ситуация, когда плохой классификатор будет обладать высокой точностью. Это может произойти в тех случаях, когда объекты в классах распределены неравномерно («перекошенные» классы, skewed classes), как пример рассмотрим классификацию по двум классам при том, что объектов первого класса — 90%, а второго — 10%,

тогда если всегда относить объект к первому классу, то точность классификатора будет равна 90%, хотя он никогда не будет определять объекты второго класса, что может быть важно в решаемой задаче. [17]

Для того, чтобы получить более адекватную оценку качества классификатора, стоит также рассматривать другие метрики, наиболее часто используются точность (precision), полнота (recall) и F-мера, которая является средним гармоническим точности (precision) и полноты [17]:

- Точность: $\mathrm{Pr}_i = \frac{\mathrm{TP}_i}{(\mathrm{TP}_i + \mathrm{FP}_i)}$;
- Полнота: $\mathrm{R}_i = \frac{\mathrm{TP}_i}{(\mathrm{TP}_i + \mathrm{FN}_i)}$;
- F-мера: $\mathrm{F1}_i = 2 \cdot \frac{\mathrm{Pr}_i \cdot R_i}{(\mathrm{Pr}_i + R_i)} = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \frac{1}{2}(\mathrm{FP}_i + \mathrm{FN}_i)}$;

Чтобы оценить производительность по каждому классу в наборе данных, рассчитывают общие показатели для каждого класса, такие как точность (precision), полнота и F-мера.

Эти метрики особенно полезны, когда оценки по классам распределены неравномерно (например, большинство экземпляров принадлежат одному классу). В таких случаях точность (accuracy) может вводить в заблуждение, так как при отнесении значения к доминирующему классу можно достичь относительно высокой общей точности, но очень низкой точности (precision) или полноты для других классов.

Точность (precision) определяется как доля правильных оценок для определенного класса, тогда как полнота — это доля экземпляров класса, которые были правильно определены. Можно отметить, что между этими двумя метриками существует очевидная связь. Когда классификатор пытается определить все как один класс, скажем, класс a, по большей части, он достигнет высокой полноты для a (будет идентифицировано большинство экземпляров этого класса). Однако экземпляры других классов, скорее всего, будут неверно определены как a, что

приведет к снижению точности (precision) для a. В дополнение к точности и полноте, также часто приводится F-мера. F-мера определяется как среднее гармоническое (или средневзвешенное значение) точности и полноты.

Метрики для каждого класса могут быть усреднены по всем классам, что приводит к макро-усредненным точности, полноте и F-мере (берется среднее арифметическое): $\mathrm{Pr} = \frac{\sum_{i=1}^{m} \mathrm{Pr}_i}{m}$, $\mathrm{R} = \frac{\sum_{i=1}^{m} \mathrm{R}_i}{m}$, $\mathrm{F1} = \frac{\sum_{i=1}^{m} \mathrm{F1}_i}{m}$.

Когда экземпляры не распределены равномерно по классам, полезно оценить работу модели по отношению к одному классу за раз, прежде чем усреднить метрики.

Одной из самых показательных метрик для данного исследования является средняя точность (average accuracy) — это доля правильно классифицированных экземпляров в матрицах ошибок для каждого класса, чтобы посчитать среднюю точность строятся матрицы ошибок для каждого класса отдельно, подсчитывается сумма их диагоналей (значения, верно отнесенные к классу и верно не включенные в класс) и делится на количество классов и общее число значений.

$\mathrm{Avg.Acc} = \frac{\sum_{i=1}^{m} (\mathrm{TP}_i + \mathrm{TN}_i)}{N \cdot m}$, где $\mathrm{TP}_i + \mathrm{TN}_i$ — доля правильно определенных значений для каждого класса i.

Если при работе классификатора отнесение экземпляров к соседним классам не является критической ошибкой, также можно рассматривать смежную точность (adjacent accuracy), отношение правильно классифицированных и отнесенных к соседним классам значений к общему числу.

Для того, чтобы оценить качество модели нужно иметь некоторое тестовое множество, которое содержит достоверную информацию по данным, оцениваемым моделью.

Мера каппа Коэна [138] обычно применяется для измерения эффективности дискретизаторов в терминах обобщения уровня классификации. Это альтернатива

для точности (accuracy). Изначальная цель этой меры было измерение степень согласия или несогласия между людьми, наблюдающими одно явление.

Каппу Коэна можно посчитать, используя матрицу ошибок: $\kappa = \frac{N\sum_{i=1}^{m} y_{ii} - \sum_{i=1}^{m} y_{i.} y_{.i}}{N^2 - \sum_{i=1}^{m} y_{i.} y_{.i}}$, где y_{ii} — количество ячеек на главной диагонали, N число экземпляров выборки, m — число классов и $y_{i.}$, $y_{.i}$ — столбцы и строчки матрицы. Значения каппы Коэна расположены в интервале от -1 (полное несогласие) до 1 (полное согласие) и проходят через 0 (рандомная классификация). Так как это скаляр, она менее выразительна, чем кривые ROC, применительно к бинарной классификации, однако для большего количества классов, каппа полезная и простая мера для измерения точности классификатора, компенсирующая случайный успех.

Одними из самых популярных информационных критериев являются AIC и BIC.

Информационный критерий Акаике AIC вычисляется следующим образом [84]: $\mathrm{AIC} = -2l + 2k$, где l — значение логарифмической функции правдоподобия построенной модели, а k — количество использованных (оцененных) параметров.

Модель с наименьшим AIC является оптимальной [244].

Использование AIC подразумевает, что каждое наблюдение дает новую, независимую информацию относительно базовой модели, которая может стать нереалистичной, когда исследуемая выборка увеличивается в размерах.

Поэтому был предложен критерий BIC [212]: $\mathrm{BIC} = -2l + k \cdot lnn$, где n — размер выборки.

Данный критерий налагает больший штраф на увеличение количества параметров по сравнению с AIC. BIC имеет свои корни в байесовской статистике, но парадоксально, что он почти всегда применяется в контексте выбора частотной (т. е. не байесовской) модели. Различие между BIC двух моделей можно рассматривать как асимптотическое приближение к логарифму фактора Байеса

этих моделей [114], который дает указание на поддержку одной модели против другой. Привлекательным свойством BIC является то, что он непротиворечив: BIC выбирает правильную модель с вероятностью, которая стремится к 1 с увеличением размера выборки [114]. Это свойство согласованности подразумевает, что истинная модель находится среди множества моделей для выбора [244].

На основе теоретических различий между AIC и BIC [244] были представлены некоторые эмпирические правила для выбора между ними. В случае, когда человек заинтересован в идентификации истинной модели (т. е. работает при нулевой/одной функции потерь) и что истинная модель имеет фиксированное и конечное число параметров, BIC предпочтительнее (в асимптотической ситуации, из-за ее свойства согласованности, упомянутого выше). Однако, когда истинная модель не входит в набор выбора и/или, когда эта истинная модель слишком сложна для параметрического моделирования (например, неизвестная, сильно нелинейная модель), AIC будет минимизировать (асимптотически) среднеквадратичная ошибку оценивания эффективнее, чем BIC.

Кросс-валидация — процедура эмпирического оценивания моделей [98]. Наиболее часто используется кросс-валидация по k классам. Это происходит следующим образом: набор данных делится на k классов, затем на $k-1$ классах проводится обучение модели, а на оставшемся классе тестирование модели, эта процедура повторяется k раз, на каждом шаге для проверки выбирается новый блок.

2.4 ОПИСАНИЕ ИССЛЕДОВАНИЯ

Формальная постановка задачи выглядит следующим образом. **Дано:** исходная модель оценивания интенсивности пуассоновского процесса на основе БСД $M=\left(G(V,L),\mathbf{P}\right)$ (см. раздел 2.2), сведения респондентов о последних эпизодах и рекордных интервалах пуассоновского процесса, выступающего математической моделью эпизодического поведения, то есть поступает свидетельство $E=\left\{t_{01},t_{12},t_{23},t_{\min},t_{\max}\right\}$, в котором данные могут быть ошибочны.

Требуется: построить такие модели (определить структуру БСД и тензоры условной вероятности), что их средняя точность классификации статистически выше, чем у исходной модели. Задача моделей оценивания интенсивности пуассоновского процесса правильно отнести величину λ к соответствующему классу дискретизации ($\lambda^{(1)}$, $\lambda^{(2)}$, …, $\lambda^{(m)}$, где m — это количество классов), а $r^0 = \max\limits_{i=1,\ldots,m} \mathrm{P}(\lambda_i \mid E)$ — это оценки величины λ исходной модели с учетом поступившего свидетельства E, такие оценки можно представить в виде матрицы ошибок (см. раздел 2.3). Требуется предложить модель $M^* = \left(G^*(V^*, L^*), \mathbf{P}^*\right)$, оценки $r^* = \max\limits_{i=1,\ldots,m} \mathrm{P}(\lambda_i \mid E)$ которой будут такими, что средняя точность r^* будет больше средней точности r^0.

Возможное решение данной задачи состоит в том, чтобы предложить подходы к обработке неопределенности, связанной с данными респондентов, в частности предложить модель $M^* = \left(G^*(V^*, L^*), \mathbf{P}^*\right)$, где $V^* = \{t_{01}, t_{12}, t_{23}, t_{\min}, t_{\max}, \lambda, n, U\}$, где U — это добавочные узлы БСД, описывающие неопределенность данных.

При оценивании интенсивности пуассоновского процесса как модели эпизодического поведения на основе моделей БСД по данным респондентов в работе рассматривается неопределенность следующих видов:

1) Несогласованность ответов респондентов.
2) Некорректность ответов респондентов.
3) Некорректность задания момента окончания исследования (момент сбора информации).

Таким образом задача диссертационного исследования заключается в том, чтобы предложить методы и алгоритмы обработки указанных видов неопределенности, а также разработать архитектуру и прототип комплекса программ, реализующих работу с предложенными методами и алгоритмами.

Чтобы сравнить показатели качества работы модели оценивания интенсивности пуассоновского процесса, предложенной ранее (раздел 2.2) используются метрики качества, описанные в разделе 2.3. Социальные сети (публикация постов рассматривается как пуассоновский процесс) используются в качестве источника данных для тестирования моделей, также используются синтетические данные.

В ходе работы с задачами выполняются следующие этапы [13]:

1 этап включает в себя постановку задачи, идентификацию рассматриваемых типов неопределенности в контексте существующей модели оценивания интенсивности поведения по ограниченному объему данных.

2 этап включает экспертное задание структуры БСД, основанное на модели оценивания интенсивности пуассоновского процесса, предложенной ранее, а также на добавлении новых структурных связей и переменных для обработки возникающих видов неопределенности.

3 этап включает сбор данных из различных источников, их описание и дискретизацию для использования в построенной модели сети, а также синтез данных.

4 этап заключается в проведении байесовского вывода с использованием собранных данных для исходной модели оценивания интенсивности пуассоновского процесса на основе БСД и для предлагаемой модели.

5 этап заключается в вычислении метрик качества предлагаемой модели по сравнению с исходной.

6 этап заключается в вычислении значений информационных критериев для исходной и предлагаемой модели на основании полученных данных и сравнении их между собой.

7 этап реализация прототипа комплекса программ для работы с предложенными методами и моделями.

ВЫВОДЫ ПО ГЛАВЕ 2

В первом разделе представлены вводные понятия теории БСД, инструмента, лежащего в основе моделей оценивания интенсивности пуассоновского процесса. Во втором разделе представлен подход к оцениванию интенсивности пуассоновского процесса с использованием БСД, разработанный А.В. Суворовой, А.Л. Тулупьевым, Т.В. Тулупьевой и соавторами, который лежит в основе данного диссертационного исследования. Отмечено, что этот подход не обладает достаточным функционалом для учета несогласованности и некорректности ответов респондентов, а также некорректности задания момента окончания исследования. Задача оценивания интенсивности пуассоновского процесса по ограниченному объему данных в случае использования байесовских сетей доверия является задачей классификации, поэтому качество алгоритмов ее решения может быть оценено посредством повышения метрик классификации. В третьем разделе рассматриваются возможные метрики оценивания качества построенных моделей: точность (accuracy), средняя точность, точность (precision) и полнота, F-мера, мера каппа Коэна, AIC и BIC. В четвертом разделе приведено описание диссертационного исследования.

Данная глава не включает результаты данного диссертационного исследования, а предназначена для введения единой системы обозначений и описания подхода к оцениванию интенсивности пуассоновского процесса, в развитии которого заключается суть диссертационного исследования.

ГЛАВА 3. МЕТОДЫ И АЛГОРИТМЫ ОБРАБОТКИ НЕОПРЕДЕЛЕННОСТИ ПРИ ОЦЕНИВАНИИ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА

Глава представляет собой описание полученных соискателем теоретических результатов. В ней описаны предложенные метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанной на БСД; алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида; метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений. Рассмотрена модель оценивания интенсивности пуассоновского процесса, расширенная скрытыми переменными, как истинными данными о последних эпизодах и рекордных интервалах пуассоновского процесса, со структурой, обученной на синтетических данных. Описаны модели оценивания интенсивности постинга в социальных сетях, расширенные за счет объективных данных о пользователе. Также рассмотрены возможные варианты дискретизации непрерывных величин, входящих в модели. Изложение материала основано на публикациях [28, 35–56, 230–237].

3.1 ОЦЕНИВАНИЕ СОГЛАСОВАННОСТИ ИНФОРМАЦИИ ОБ ИНТЕРВАЛАХ ПУАССОНОВСКОГО ПРОЦЕССА

Для того, чтобы определить, насколько согласованы между собой данные, характеризующие длины интервалов между окончанием периода исследования, эпизодами пуассоновского процесса и длины рекордных интервалов, то есть нет ли в них внутренних противоречий, предлагается метод оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанный на расширении

модели узлами, которые отражают согласованность информации о последних эпизодах и рекордных интервалах (рис. 3.1). Инструмент оценивания согласованности данных об интервалах пуассоновского процесса особенно полезен в тех случаях, когда собираются данные от респондентов, например, данные об эпизодах какого-то определенного типа поведения этих респондентов.

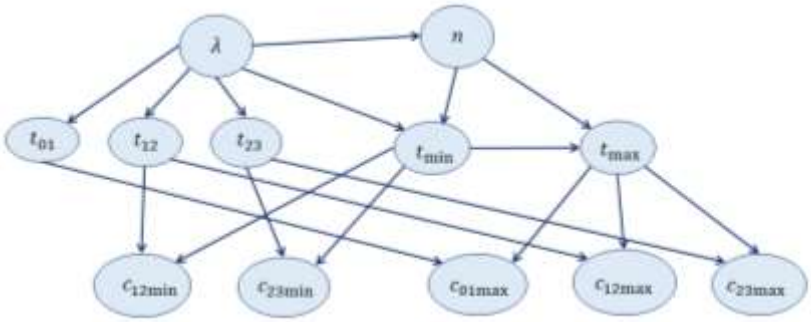

Рисунок 3.1 — Расширенная модель оценивания интенсивности пуассоновского процесса с диагностикой согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса [38]

Вероятность согласованности означиваний случайных величин t_{12} и t_{23}, которые соответствуют интервалам между последним и предпоследним эпизодами пуассоновского процесса и предпоследним и предпредпоследним, со значением случайной величины $t_{\min}$, которая соответствует длине минимального интервала между эпизодами пуассоновского процесса за исследуемый период, рассчитывается следующим образом: если данные противоречат друг другу (интервал меньше минимального), то она равна 0, если данные согласованы — 1. Согласованность t_{01} и $t_{\min}$ не рассматривается в связи с тем, что t_{01} соответствует времени , разделяющему факт получения данных, и последний эпизод. Этот промежуток не есть интервал между эпизодами пуассоновского процесса.

$c_{t_{ij},\min}$ $(i=1,2;\quad j=i+1)$, определяющая оценку согласованности данных, может иметь одно из следующих трех значений: свидетельства t_{ij} и $t_{\min}$ согласованы (обозначим $c^{+}_{t_{ij},\min}$), не согласованы ($c^{-}_{t_{ij},\min}$), и согласованность не

определена ($c^{?}_{t_{ij,\min}}$). Последнее значение $c_{t_{ij,\min}}$ принимает, если значения t_{ij} и $t_{\min}$ относятся к одному интервалу, то есть $t_{ij} \in [a;b)$ и $t_{\min} \in [a;b)$, в этом случае не получится определить, меньше ли значение $t_{\min}$ по отношению к t_{ij}.

Тензоры условной вероятности $P\left(c^{(s)}_{t_{ij,\min}} \mid t_{ij}, t_{\min}\right)$ задаются следующим образом:

$$P\left(c^{(s)}_{t_{ij,\min}} \mid t_{ij}, t_{\min}\right) = \begin{cases} \alpha^{(s)}, & t_{ij} > t_{\min}; \\ \beta^{(s)}, & t_{ij} < t_{\min}; \\ 1-\alpha^{(s)}-\beta^{(s)}, & t_{ij} = t_{\min}; \end{cases}$$

где $s \in \{+,-,?\}$, $\alpha^{(s)}, \beta^{(s)} \in [0;1]$, $\alpha^{(s)} + \beta^{(s)} \le 1$, $\sum\limits_{s\in\{+,-,?\}} \alpha^{(s)} = 1$, $\sum\limits_{s\in\{+,-,?\}} \beta^{(s)} = 1$.

Разберем следующий частный случай: $\alpha^{(s)}, \beta^{(s)} \in \{0,1\}$, $s \in \{+,-,?\}$. При этом вероятность согласованности данных будет равна нулю при противоречиях в них, и единице при отсутствии противоречий.

t_{ij} разделим на непересекающиеся интервалы $t^{(1)}_{ij}, \ldots, t^{(n)}_{ij}$, и $t_{\min}$ разобьем на такие же интервалы $t^{(1)}_{\min}, \ldots, t^{(n)}_{\min}$, то есть $t^{(i)}_{ij} = t^{(i)}_{\min}$ для всех $i \in [1;n]$. Тогда распределение $P\left(c^{+}_{t_{ijmin}} \mid t_{ij}, t_{min}\right)$ будет задаваться так:

$$\begin{cases} P\left(c^{+}_{t_{ijmin}} \mid t^{(s)}_{ij}, t^{(k)}_{min}\right) = 1, \ при\, k < s \\ P\left(c^{+}_{t_{ijmin}} \mid t^{(s)}_{ij}, t^{(k)}_{min}\right) = 0, \ при\, k \ge s \end{cases},$$

распределение $P\left(c^{-}_{t_{ijmin}} \mid t_{ij}, t_{min}\right)$ — $\begin{cases} P\left(c^{-}_{t_{ijmin}} \mid t^{(s)}_{ij}, t^{(k)}_{min}\right) = 1, \ при\, k > s \\ P\left(c^{-}_{t_{ijmin}} \mid t^{(s)}_{ij}, t^{(k)}_{min}\right) = 0, \ при\, k \le s \end{cases}$,

а распределение $P\left(c^{?}_{t_{ijmin}} \mid t_{ij}, t_{min}\right)$ — $\begin{cases} P\left(c^{?}_{t_{ijmin}} \mid t^{(s)}_{ij}, t^{(k)}_{min}\right) = 1, \ при\, k = s \\ P\left(c^{?}_{t_{ijmin}} \mid t^{(s)}_{ij}, t^{(k)}_{min}\right) = 0, \ при\, k \ne s \end{cases}$.

Аналогичным образом рассматривается согласованность значений случайных величин t_{ij} $(i = 0,1,2;\ j = i+1)$, со значением случайной величины $t_{\max}$

.

$c_{t_{ij,\max}}$, определяющая оценку согласованности данных, может иметь одно из следующих трех значений: свидетельства t_{ij} и $t_{\max}$ согласованы (обозначим $c^{+}_{t_{ij,\max}}$), не согласованы ($c^{-}_{t_{ij,\max}}$), и согласованность не определена ($c^{?}_{t_{ij,\max}}$).

Тензоры условной вероятности $P\left(c^{(s)}_{t_{ij,\max}} \mid t_{ij}, t_{\max}\right)$ задаются следующим образом:

$$P\left(c^{(s)}_{t_{ij,\max}} \mid t_{ij}, t_{\max}\right) = \begin{cases} \alpha^{(s)}, & t_{ij} < t_{\max}; \\ \beta^{(s)}, & t_{ij} > t_{\max}; \\ 1-\alpha^{(s)}-\beta^{(s)}, & t_{ij} = t_{\min}; \end{cases}$$

где $s \in \{+,-,?\}$, $\alpha^{(s)}, \beta^{(s)} \in [0;1]$, $\alpha^{(s)} + \beta^{(s)} \le 1$, $\sum \alpha = 1, \sum \beta = 1$.

Также для получения общей оценки можно добавить к этой модели вершину g (рис. 3.2), которая характеризует оценку общей согласованности данных. Для простоты обозначим $c = \left(c_{t_{12,\min}}, c_{t_{23,\min}}, c_{t_{01,\max}}, c_{t_{12,\max}}, c_{t_{23,\max}}\right)$, тогда $P(g^{+} \mid c) = \dfrac{\sum c^{+}}{\sum c}$, $P(g \mid c) = \dfrac{\sum c^{-}}{\sum c}$ и $P(g^{?} \mid c) = \dfrac{\sum c^{?}}{\sum c}$.

Отметим, что α и β являются порогами согласованности, определение которых зависит от специалистов конкретной области, использующих предложенный метод для исследований различных видов процессов.

Рассмотрим алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса на примере узла $c_{t_{12,\min}}$, остальные рассматриваются аналогичным образом. Пусть t_{12} разбита на следующие дизъюнктные промежутки: $t^{(1)} = (0;0.1)$, $t^{(2)} = [0.1;1)$, $t^{(3)} = [1;7)$, $t^{(4)} = [7;30)$, $t^{(5)} = [30;180)$, $t^{(6)} = [180;+\infty)$; для $t_{\min}$ возьмем то же самое разбиение $t^{(1)},\ldots,t^{(6)}$. Пороги согласованности заданы следующим образом: $\alpha^{+} = \beta^{-} = 1$, $\alpha^{-} = \alpha^{?} = \beta^{+} = \beta^{?} = 0$, то есть вероятность согласованности информации равна 0, если данные, в которых она содержится, противоречат друг другу, и 1, если

противоречий нет. Если поступили свидетельства $t_{12}^{(4)}$ и $t_{\min}^{(1)}$, то есть длина минимального интервала меньше, чем длина интервала между последним и предпоследним эпизодом, то информация, согласована. Если поступили свидетельства $t_{12}^{(5)}$ и $t_{\min}^{(6)}$, то есть длина минимального интервала больше, чем длина интервала между последним и предпоследним эпизодом, то информация не согласована. Если поступили свидетельства $t_{12}^{(4)}$ и $t_{\min}^{(4)}$, то есть длины минимального интервала и интервала между последним и предпоследним эпизодом относятся к одному промежутку дискретизации, то оценка согласованности информации, не определена.

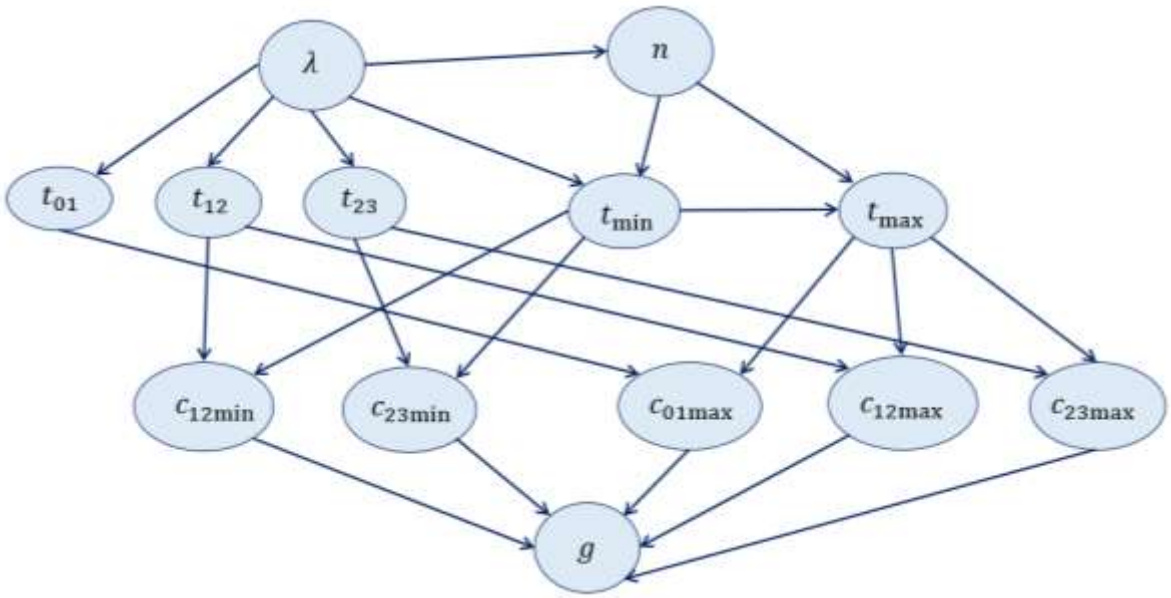

Рисунок 3.2 — Расширенная модель оценивания интенсивности пуассоновского процесса с общей оценкой согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса

Алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанной на БСД, реализованный в модуле прототипа комплекса программ, представлен на рис. 3.3 (здесь и далее для алгоритмов использована нотация BPMN).

Рисунок 3.3 — Схема алгоритма оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности

3.2 ИСПОЛЬЗОВАНИЕ СКРЫТЫХ ПЕРЕМЕННЫХ ДЛЯ МОДЕЛИРОВАНИЯ ИСТИННОЙ ИНФОРМАЦИИ ОБ ЭПИЗОДАХ ПУАССОНОВСКОГО ПРОЦЕССА

Одним из недостатков использования исходной модели является то, что она не учитывает вероятность неточности, или умышленного искажения данных, представленных респондентами в тех случаях, когда исследуются такие процессы как поведение определенного типа или другие процессы, данные о которых можно получить посредством опроса респондентов. Иногда ответ о поведении является попыткой представить его более социально одобряемым, или напротив, социально-неодобряемым. Также возможны ошибки и неточности в ответах, связанные с работой человеческой памяти.

Чтобы повысить качество оценки интенсивности пуассоновского процесса, за счет обработки такой неопределенности, предлагается алгоритм, учитывающий возможную некорректность информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида. Этот алгоритм основан на добавлении в модель скрытых переменных, содержащих истинную информацию об эпизодах пуассоновского процесса.

В модель оценивания интенсивности пуассоновского процесса добавлены вершины t_{01}^0, t_{12}^0, t_{23}^0, $t_{\min}^0$ и $t_{\max}^0$ (см. рис. 3.4), представляющие интервалы

пуассоновского процесса, полученные из ответов респондентов (то же самое, что и t_{01}, t_{12}, t_{23}, $t_{\min}$ и $t_{\max}$ в исходной модели).

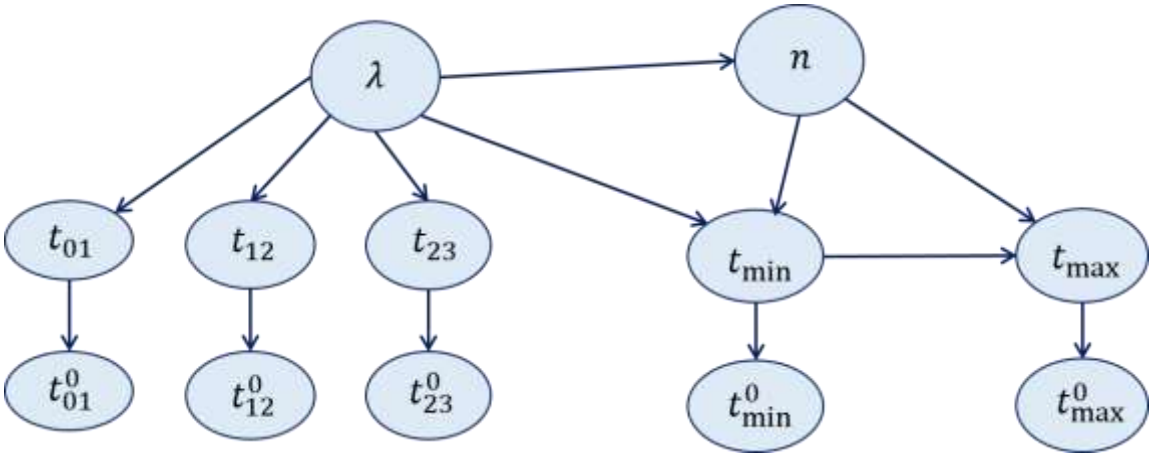

Рисунок 3.4 — Модель оценивания интенсивности пуассоновского процесса со скрытыми переменными [235]

А t_{01}, t_{12}, t_{23}, $t_{\min}$ и $t_{\max}$ теперь — это скрытые переменные, описывающие интервалы пуассоновского процесса между эпизодами, действительно случившимися. Действительные интервалы остаются неизвестными, так как источник данных — ответы респондентов, предполагающие возможную неточность информации. Искажения могут быть как случайными, так и намеренными, обусловленными желанием сформировать определенное мнение о себе в восприятии собеседника.

Тензоры условной вероятности определяются следующим образом:

$$p(t_{12}^0 \mid t_{12}) = \begin{cases} c, \mid t_{12}^0 - t_{12} \mid = 0 \\ c \cdot q, \mid t_{12}^0 - t_{12} \mid = 1 \\ c \cdot q^2, \mid t_{12}^0 - t_{12} \mid = 2 \\ \dots \\ c \cdot q^n, \mid t_{12}^0 - t_{12} \mid = n \end{cases},$$ где $q \in [0;1)$, а c — это нормализирующая константа.

Также тензоры условной вероятности можно получить, обучив параметры модели на статистических данных.

Рассмотрим результаты работы модели на синтетических данных (см. раздел 4.4.1) при разных значениях q. Для величины, характеризующей интенсивность пуассоновского процесса, была взята дискретизация интенсивности на следующие интервалы: $[0;0.01)$, $[0.01;0.05)$, $[0.05;0.1)$, $[0.1;1)$, и $[1;\infty)$. Было синтезировано 4000 записей, после удаления записей без эпизодов поведения осталось 3972 записи с известными интенсивностями.

При $q = 0$ апостериорное распределение λ показано на рис. 3.5.

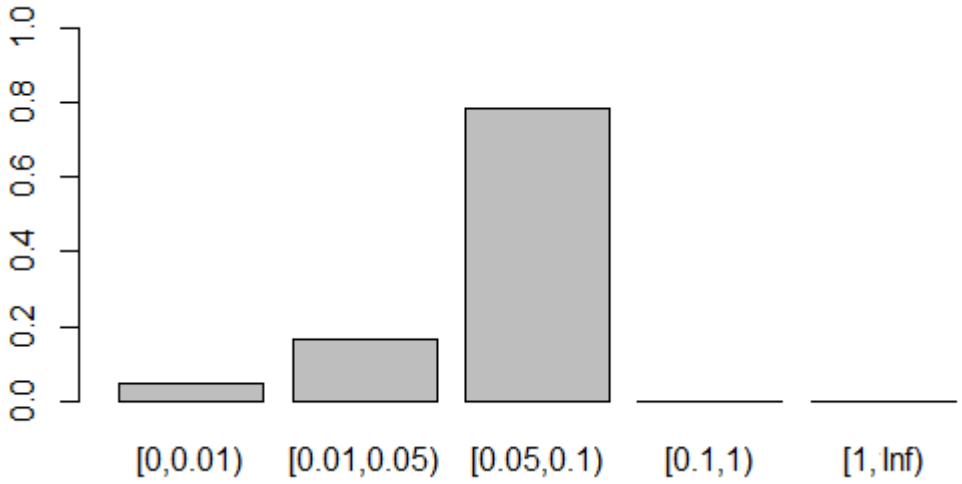

Рисунок 3.5 — Апостериорное распределение интенсивности пуассоновского процесса ($q = 0$)

Сравним эти результаты с заранее синтезированными величинами интенсивности пуассоновского процесса. Матрица ошибок модели оценивания интенсивности пуассоновского процесса со скрытыми переменными представлена таблицей 3.1.

Таблица 3.1 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса со скрытыми переменными ($q = 0$)

	[0;0.01)	[0.01;0.05)	[0.05;0.1)	[0.1;1)	[10;∞)
[0;0.01)	6	1	0	0	0
[0.01;0.05)	66	144	15	0	0
[0.05;0.1)	9	235	276	0	0
[0.1;1)	0	296	2864	0	0
[1;∞)	0	51	9	0	0

Как видно из таблицы точность (acc.) в данном случае не высока (0.107), но тут корректно рассматривать среднюю точность (avg. acc.), которая составляет 0.702. В таблице 3.2 приведены метрики качества классификации модели оценивания интенсивности пуассоновского процесса со скрытыми переменными по классам при $q = 0$.

Таблица 3.2 — Метрики качества классификации модели оценивания интенсивности пуассоновского процесса со скрытыми переменными ($q = 0$)

	точность	полнота	F1
[0;0.01)	0.074	0.857	0.136
[0.01;0.05)	0.198	0.64	0.303
[0.05;0.1)	0.087	0.531	0.15
[0.1;∞)	NaN	0	NaN

При $q = 0.3$ апостериорное распределение λ представлено на рис. 3.6, а матрица ошибок приведена в таблице 3.3, метрики качества приведены в таблице 3.4. Средняя точность равна 0.723.

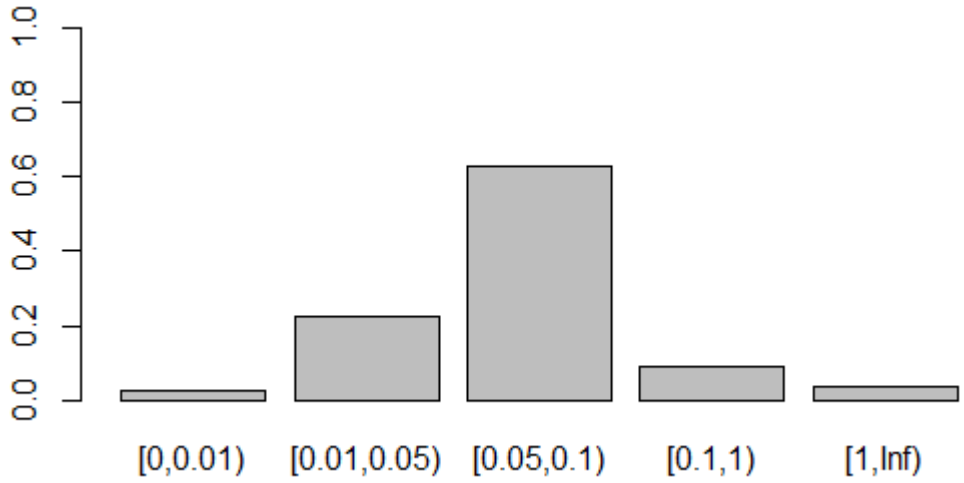

Рисунок 3.6 — Апостериорное распределение интенсивности интенсивности пуассоновского процесса ($q = 0.3$)

Таблица 3.3 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса со скрытыми переменными ($q = 0.3$)

	[0;0.01)	[0.01;0.05)	[0.05;0.1)	[0.1;1)	[1;∞)

[0;0.01)	5	2	0	0	0
[0.01;0.05)	62	144	19	0	0
[0.05;0.1)	3	227	290	0	0
[0.1;1)	0	83	2817	218	42
[1;∞)	0	0	7	44	9

Таблица 3.4 — Метрики качества классификации модели оценивания интенсивности пуассоновского процесса со скрытыми переменными ($q = 0.3$)

	точность	полнота	F1
[0;0.01)	0.071	0.714	0.13
[0.01;0.05)	0.316	0.64	0.423
[0.05;0.1)	0.093	0.558	0.159
[0.1;1)	0.832	0.069	0.127
[1;∞)	0.176	0.15	0.162

При $q = 0.6$ апостериорное распределение λ представлено на рис. 3.7, а матрица ошибок в таблице 3.5, метрики качества приведены в таблице 3.6. Средняя точность равна 0.721.

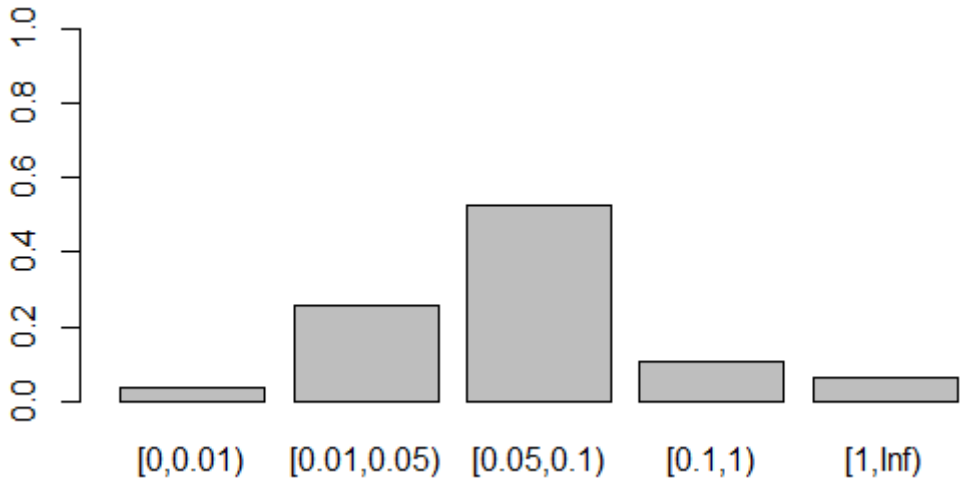

Рисунок 3.7 — Апостериорное распределение интенсивности интенсивности пуассоновского процесса ($q = 0.6$)

Таблица 3.5 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса со скрытыми переменными ($q = 0.6$)

	[0;0.01)	[0.01;0.05)	[0.05;0.1)	[0.1;1)	[1;∞)
[0;0.01)	5	2	0	0	0
[0.01;0.05)	40	145	40	0	0
[0.05;0.1)	1	178	341	0	0
[0.1;1)	0	58	2842	123	137
[1;∞)	0	0	7	21	32

Таблица 3.6 — Метрики качества классификации модели оценивания интенсивности пуассоновского процесса со скрытыми переменными ($q = 0.6$)

	точность	полнота	F1
[0;0.01)	0.109	0.714	0.189
[0.01;0.05)	0.379	0.644	0.477
[0.05;0.1)	0.106	0.656	0.182
[0.1;1)	0.854	0.039	0.074
[1;∞)	0.189	0.533	0.279

При $q = 0.9$ апостериорное распределение λ представлено на рис. 3.8, а матрица ошибок в таблице 3.7, метрики качества приведены в таблице 3.8. Средняя точность равна 0.717.

Таблица 3.7 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса со скрытыми переменными ($q = 0.9$)

	[0;0.01)	[0.01;0.05)	[0.05;0.1)	[0.1;1)	[1;∞)
[0;0.01)	4	1	2	0	0
[0.01;0.05)	3	151	71	0	0
[0.05;0.1)	0	81	439	0	0
[0.1;1)	0	9	3151	0	0
[1;∞)	0	0	60	0	0

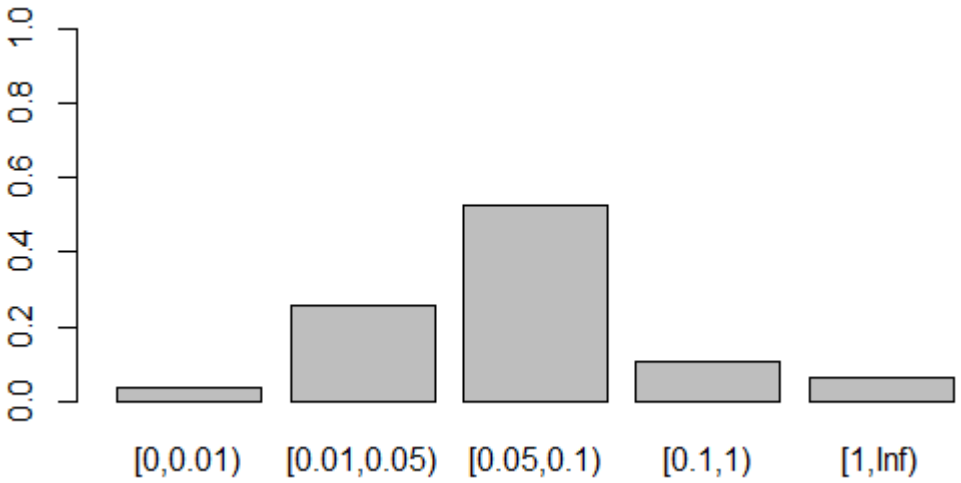

Рисунок 3.8 — Апостериорное распределение интенсивности интенсивности пуассоновского процесса ($q = 0.9$)

Таблица 3.8 — Метрики качества классификации модели оценивания интенсивности пуассоновского процесса со скрытыми переменными ($q = 0.9$)

	точность	полнота	F1
[0;0.01)	0.571	0.571	0.571
[0.01;0.05)	0.624	0.671	0.647
[0.05;0.1)	0.118	0.844	0.207
[0.1;∞)	NaN	0	NaN

Сравним эти результаты с работой исходной модели (без скрытых переменных).

Апостериорное распределение интенсивности показано на рис. 3.9, а матрица ошибок в таблице 3.9. Средняя точность равна 0.702.

Таблица 3.9 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса (исходная модель)

	[0;0.01)	[0.01;0.05)	[0.05;0.1)	[0.1;1)	[1;∞)
[0;0.01)	6	1	0	0	0
[0.01;0.05)	66	144	15	0	0
[0.05;0.1)	9	235	276	0	0
[0.1;1)	0	296	2864	0	0

[1;∞)	0	51	9	0	0

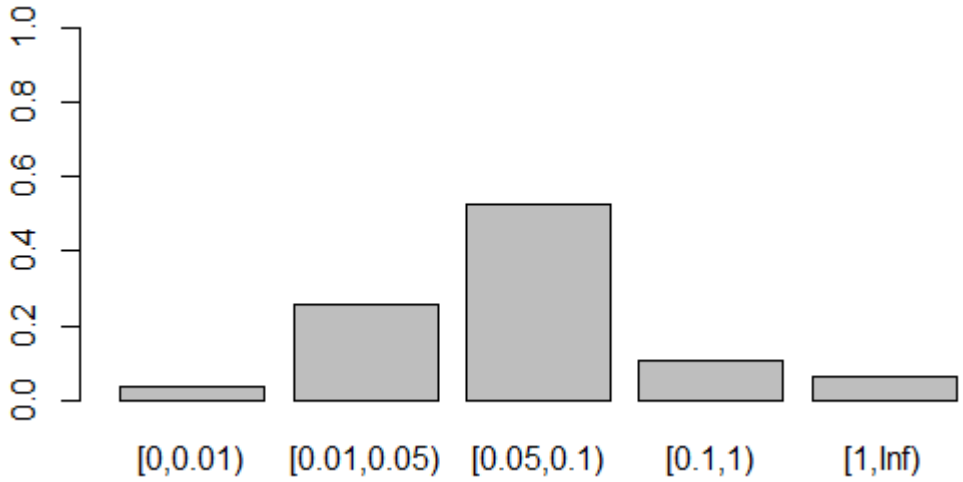

Рисунок 3.9 — Апостериорное распределение интенсивности интенсивности пуассоновского процесса (исходная модель)

Таблица 3.10 — Метрики качества классификации модели оценивания интенсивности пуассоновского процесса (исходная модель)

	точность	полнота	F1
[0;0.01)	0.074	0.857	0.136
[0.01;0.05)	0.198	0.64	0.303
[0.05;0.1)	0.087	0.531	0.15
[0.1; ∞)	NaN	0	NaN

Как видно, добавление в модель скрытых переменных способно повысить точность результатов по сравнению с исходной моделью. При этом результаты работы модели могут быть улучшены за счет выбора q, в предложенных примерах лучшие показатели качества работы модели получились при $q = 0.3$.

Рассмотрим также модель оценивания интенсивности пуассоновского процесса со скрытыми переменными, обученную на данных (рис. 3.10). Для построения структуры модели по синтетическим данным был использован алгоритм «восхождения на холм» (Hill-Climbing) с мерой качества BIC.

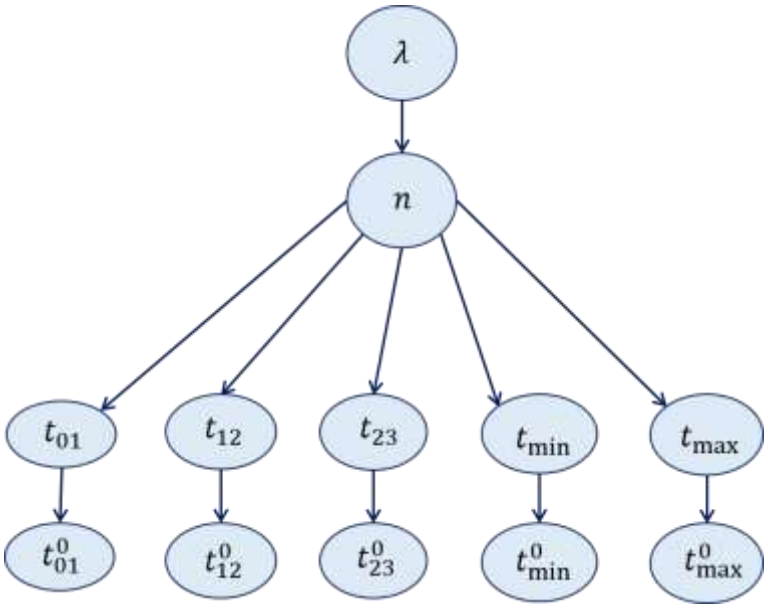

Рисунок 3.10 — Модель оценивания интенсивности пуассоновского процесса со скрытыми переменными с обученной структурой [235]

Полученную структуру можно интерпретировать следующим образом: интенсивность пуассоновского процесса λ выражается через n, количество эпизодов (за период исследования). При исследовании различных видов процессов получить точное значение n в явном виде невозможно, то есть это скрытая переменная, которая выражается посредством данных о последних эпизодах пуассоновского процесса и рекордных интервалах.

Сравнение показателей качества модели со скрытыми переменными, а также исходной модели, на синтетических данных приводится в [235]. Мера качества для модели с обученной структурой равна -55264 для байесовского информационного критерия и -53099 для меры максимального правдоподобия, для модели с экспертно заданной структурой — -70194 для байесовского информационного критерия и -53277 для меры максимального правдоподобия, значение байесовского информационного критерия для исходной модели равно -53293 и -37027 для меры максимального правдоподобия.

На тестовой выборке были получены следующие показатели: мера качества для модели с обученной структурой равна -15435 для байесовского информационного критерия и -13270 для меры максимального правдоподобия, для

модели с экспертно заданной структурой — -24068 для байесовского информационного критерия и -10379 для меры максимального правдоподобия.

Матрицы ошибок для моделей представлены в таблице 3.11 (модель оценивания интенсивности пуассоновского процесса со скрытыми переменными), таблице 3.12 (модель оценивания интенсивности пуассоновского процесса со скрытыми переменными с обученной структурой) и таблице 3.13 (исходная модель).

Таблица 3.11 — Оценки интенсивности пуассоновского процесса модели со скрытыми переменными

		Оценка интенсивности									
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$	$\lambda^{(10)}$
Значение интенсивности	$\lambda^{(1)}$	2	6	3	3	1	0	0	0	0	0
	$\lambda^{(2)}$	0	5	4	3	0	2	0	0	0	0
	$\lambda^{(3)}$	0	7	23	18	4	7	1	0	0	0
	$\lambda^{(4)}$	0	2	12	22	5	13	6	0	0	0
	$\lambda^{(5)}$	0	0	9	14	8	19	24	1	0	0
	$\lambda^{(6)}$	1	0	0	8	2	7	22	5	0	0
	$\lambda^{(7)}$	0	0	2	6	5	30	166	130	6	0
	$\lambda^{(8)}$	0	0	1	0	0	3	48	464	69	0
	$\lambda^{(9)}$	0	0	0	0	0	0	1	141	92	21
	$\lambda^{(10)}$	0	0	0	0	0	0	0	4	12	14

Заметим, что при данной дискретизации переменной задача оценивания интенсивности пуассоновского процесса является классификационной задачей для 10 непересекающихся классов.

В таблице 3.14 предствлено сравнение всех моделей по наиболее распространенным показателям качества классификации — точности (accuracy), средней точности (average accuracy), точности (precision) и полноте (recall). Модель со скрытыми переменными имеет лучшие показатели качества классификации по сравнению с исходной моделью и моделью со скрытыми переменными с обученной структурой.

Таблица 3.12 — Оценки интенсивности пуассоновского процесса модели со скрытыми переменными с обученной структурой

		Оценка интенсивности									
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$	$\lambda^{(10)}$
Значение интенсивности	$\lambda^{(1)}$	2	6	5	2	0	0	0	0	0	0
	$\lambda^{(2)}$	0	3	4	6	0	1	0	0	0	0
	$\lambda^{(3)}$	3	2	22	17	7	8	1	0	0	0
	$\lambda^{(4)}$	0	5	11	20	4	10	9	1	0	0
	$\lambda^{(5)}$	0	2	2	20	7	17	27	0	0	0
	$\lambda^{(6)}$	0	0	2	4	1	11	23	4	0	0
	$\lambda^{(7)}$	0	0	1	7	3	27	154	149	3	1
	$\lambda^{(8)}$	0	0	1	0	0	2	45	468	68	1
	$\lambda^{(9)}$	0	0	0	0	0	0	1	148	90	16
	$\lambda^{(10)}$	0	0	0	0	0	0	0	3	14	13

Таблица 3.13 — Оценки интенсивности пуассоновского процесса исходной модели

		Оценка интенсивности									
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$	$\lambda^{(10)}$
Значение интенсивности	$\lambda^{(1)}$	4	5	5	1	0	0	0	0	0	0
	$\lambda^{(2)}$	1	3	5	3	0	2	0	0	0	0
	$\lambda^{(3)}$	0	3	17	29	2	7	2	0	0	0
	$\lambda^{(4)}$	0	2	2	32	2	15	7	0	0	0
	$\lambda^{(5)}$	0	0	2	22	2	26	23	0	0	0
	$\lambda^{(6)}$	0	0	2	8	0	7	24	4	0	0
	$\lambda^{(7)}$	0	0	0	5	1	33	165	139	2	0
	$\lambda^{(8)}$	0	0	1	0	0	0	50	456	78	0
	$\lambda^{(9)}$	0	0	0	0	0	0	0	152	78	25
	$\lambda^{(10)}$	0	0	0	0	0	0	0	2	15	13

Алгоритм обработки возможной некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, по ограниченному

объему доступных наблюдений, использующий модель БСД со скрытыми переменными, представлен на рис. 3.11.

Таблица 3.14 — Сравнение метрик качества моделей оценивания интенсивности пуассоновского процесса исходной модели

	Точность (accuracy)	**Ср. точность**	**Точность (precision)**	**Полнота**
Исходная модель	0.524	0.905	0.422	0.348
Модель со скрытыми переменными	0.541	0.908	0.42	0.36
Модель со скрытыми переменными с обученной структурой	0.532	0.906	0.388	0.341

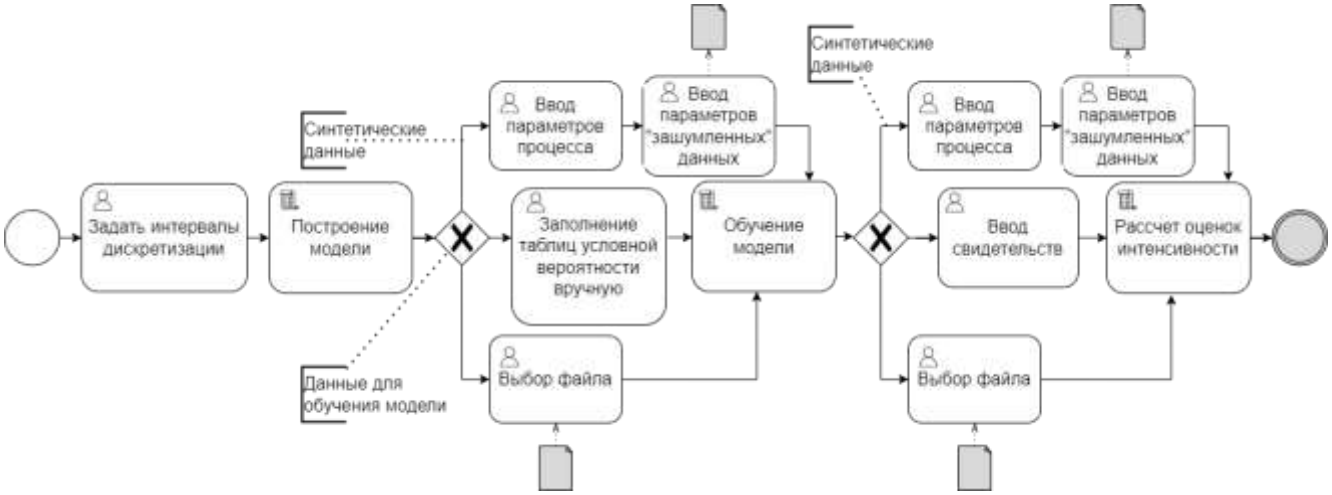

Рисунок 3.11 — Схема алгоритма обработки некорректности информации об эпизодах поведения, полученной от респондентов, при оценивании интенсивности пуассоновского процесса

3.3 ОБРАБОТКА НЕОПРЕДЕЛЕННОСТИ ЗАДАНИЯ КОНЦА ИССЛЕДУЕМОГО ПЕРИОДА ПРИ ОЦЕНИВАНИИ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА

Предлагается новый метод обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений, использующий при обучении модели сведения о значениях длин интервалов между последними тремя эпизодами, интервале между последним эпизодом и гипотетическим «следующим» и рекордных интервалах [48]. То есть в данном методе обрабатывается неопределенность, возникающая при некорректном задании окончания исследуемого периода. Промежуток времени между последним эпизодом и моментом завершения исследуемого периода отличается от интервалов между последовательными эпизодами; его представление и обработка, таким образом, требуют подходов, учитывающих указанную особенность.

Метод основан на том, что в модель оценивания интенсивности пуассоновского процесса на этапе обучения вводится вершина, характеризующая интервал между последним эпизодом пуассоновского процесса и эпизодом, произошедшим после окончания периода исследования, что позволяет повысить качество оценки интенсивности пуассоновского процесса.

Модель оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом также представляет собой БСД (рис. 3.12). В качестве исходных данных предлагаемая модель принимает сведения о продолжительности интервалов между тремя последними эпизодами поведения. Обозначим их следующим образом: t_{12}, t_{23}, $t_{\max}$ и $t_{\min}$, где t_{12} — значение длины интервала между двумя последними эпизодами пуассоновского процесса, то есть последним и предпоследним, t_{23} — значение длины интервала между предпредпоследним (третьим с конца исследуемого периода) и предпоследним эпизодом, $t_{\max}$ и $t_{\min}$ — соответственно сведения о максимальном и минимальном

значениях длин интервалов между эпизодами за исследуемый период. t_{01} — значение длины интервала между окончанием периода исследования и последним эпизодом пуассоновского процесса. В качестве исходных данных могут выступать как все перечисленные, так и любое непустое подмножество указанного перечня. Скрытая переменная t_{next} соответствует интервалу между последним эпизодом, входящим в исследуемый период, и первым эпизодом, который произойдет по окончании исследуемого периода. Например, период исследования — это 2020-ый год, последний эпизод исследуемого поведения произошел 29-го декабря, а первый эпизод по окончании исследуемого периода — 6-е января, тогда t_{next} соответствует интервалу от 29-го декабря до 6 января. Величина λ характеризует интенсивность пуассоновского процесса, скрытая переменная n — число эпизодов за исследуемый период.

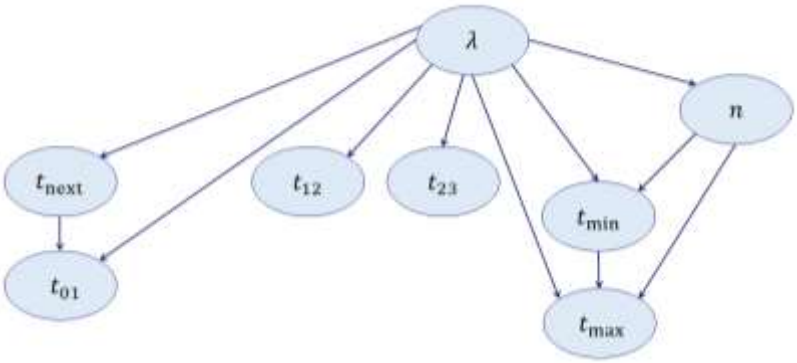

Рисунок 3.12 — Модель оценивания интенсивности пуассоновского процесса, обрабатывающая неопределенность задания конца исследуемого периода, по ограниченному объему доступных наблюдений [48]

При построении БСД используется разбиение области допустимых значений непрерывных величин на конечное число дизъюнктных интервалов. При исследовании процесса, интенсивность которого за изучаемый период может быть равна нулю, первый такой интервал рекомендуется взять таким образом, что начинается он от нуля включительно, а заканчивается значением интенсивности меньшим, чем при хотя бы одном эпизоде поведения за изучаемый период. Подробнее про дискретизацию непрерывных величин написано в разделе 3.5.

Таблицы условных вероятностей строятся на основе данных, при отсутствии достаточного количества данных для обучения можно использовать синтетические данные (см. раздел 4.4.1). Машинное обучение параметров БСД может быть проведено с использованием программы для работы с данной моделью (см. раздел 4.3, используется метод максимального правдоподобия) или других программных средств (см. раздел 1.5). Обучение параметров сети сводится к вычислению условных вероятностей для каждой пары переменных, соединенных ребром. Основная задача построенной модели — автоматизация оценивания интенсивности пуассоновского процесса. Необходимо автоматизировать получение оценки интенсивности пуассоновского процесса при возможном дефиците информации: данные могут быть неточными и неполными, а доступны только сведения о длинах интервалов между несколькими последними эпизодами, минимальном и максимальном значении длин интервалов между эпизодами. Автоматизация получения оценки интенсивности поведения обеспечена в программе для работы с моделью оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом, созданной в рамках данного диссертационного исследования (см. раздел 4.3), также для этой цели могут быть использованы другие инструменты (см. раздел 1.5). Результаты работы модели на различных данных приведены в разделе 4.5.1.

Данная модель может повысить качество оценки в тех случаях, если для обучения модели имеется набор ретроспективных данных, таким образом t_{next} на этапе обучения не будет являться скрытой переменной.

Алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений на основе БСД, реализованный в модуле прототипа комплекса программ, представлен на рис. 3.13.

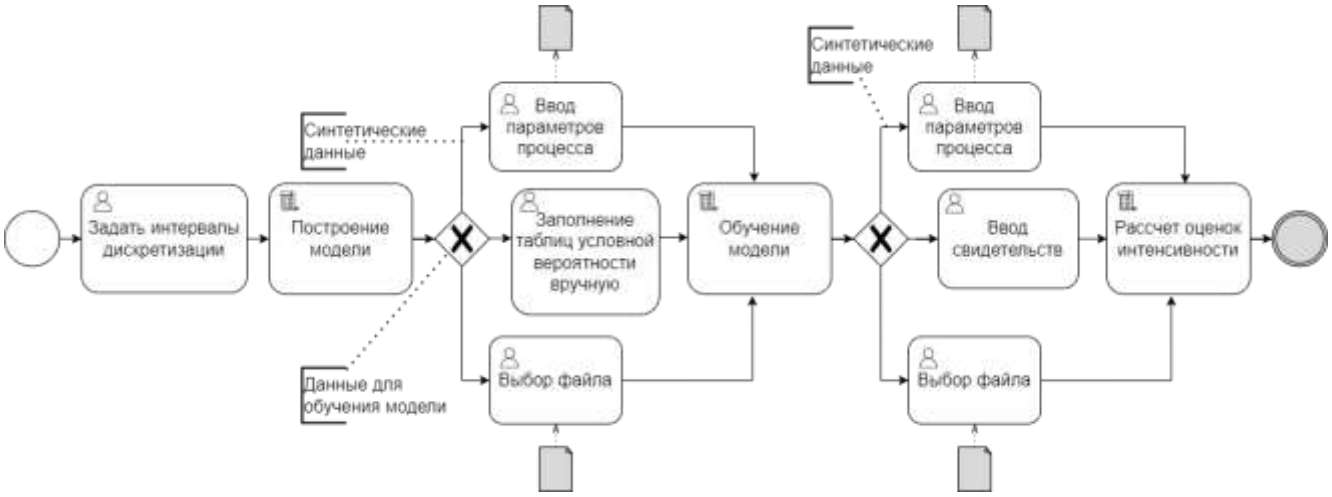

Рисунок 3.13 — Схема алгоритма обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений

3.4 ВЕРОЯТНОСТНАЯ ГРАФИЧЕСКАЯ МОДЕЛЬ ОЦЕНИВАНИЯ ИНТЕНСИВНОСТИ ПОСТИНГА В СОЦИАЛЬНОЙ СЕТИ С УЧЕТОМ ОБЪЕКТИВНЫХ ДЕТЕРМИНАНТ ПОВЕДЕНИЯ

Модель оценивания интенсивности пуассоновского процесса можно в том числе использовать и для оценивания частоты постинга в социальных сетях. Обычно из социальных сетей можно также получить информацию о поле, возрасте, количестве «друзей» в социальной сети и других характеристиках пользователя. Таким образом была поставлена задача расширить модель оценивания интенсивности пуассоновского процесса за счет этой дополнительной информации об индивиде.

На рисунке 3.14 представлена структура модели оценивания интенсивности постинга, включающая объективные данные о пользователе (о поле *sex*, возрасте *age* и числе друзей пользователя *friends count*).

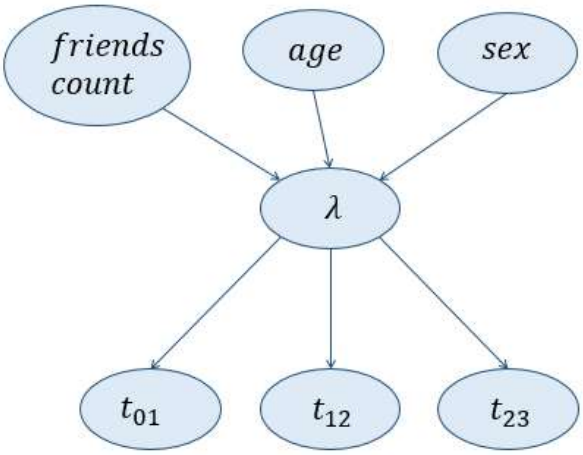

Рисунок 3.14 — Модель оценивания интенсивности постинга, включающая объективные данные о пользователе [28]

Также была рассмотрена обученная структура (рисунок 3.15) на данных тренировочной выборки с помощью алгоритма оптимизации Hill-Climbing (HC). В качестве основной меры качества был использован информационный критерий Акаике (AIC).

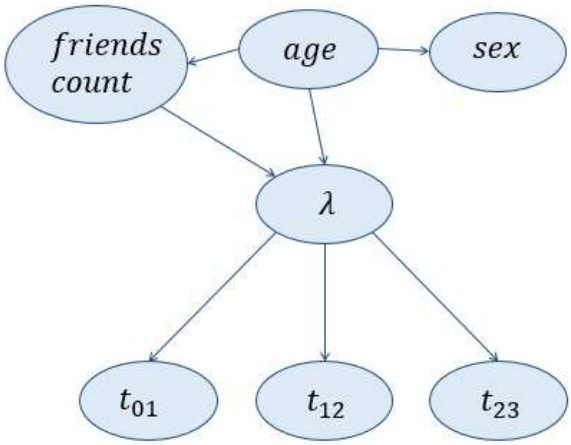

Рисунок 3.15 — Модель оценивания интенсивности постинга, включающая объективные данные о пользователе с обученной структурой [28]

Для оценивания силы взаимосвязи переменных в полученной структуре использовалась разница между правдоподобием моделей с дугой и без нее. В таблице 3.15 показана сила связей дуг для обученной структуры. Чем меньше значение, тем больше сила связи.

Таблица 3.15 — Сила связей дуг обученной структуры модели [28]

Начало дуги	Конец дуги	сила
λ	t_{01}	-299,54
λ	t_{12}	-205,381
λ	t_{23}	-190,153
friends count	λ	-24,478
age	*friends count*	-15,039
age	λ	-5,121
age	*sex*	-2,549

3.5 ДИСКРЕТИЗАЦИЯ НЕПРЕРЫВНЫХ ВЕЛИЧИН В МОДЕЛЯХ ОЦЕНИВАНИЯ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА

При построении БСД используется разбиение области допустимых значений непрерывных величин на конечное число интервалов. В данном диссертационном исследовании не рассматриваются непрерывные БСД, так как было доказано, что точный вывод возможен только для дискретных, гауссовых и условно-линейных гауссовых БСД, однако практически невозможно найти наборы данных, где переменные следуют гауссовому распределению [214]. Для работы с описанными моделями оценивания интенсивности пуассоновского процесса необходимо провести дискретизацию случайных величин $t_{i,j+1}$, $t_{\min}$ и $t_{\max}$, характеризующих длины временных интервалов между эпизодами, и случайной величины, характеризующей интенсивность пуассоновского процесса λ.

Дискретизация может проводится по размеченному и неразмеченному набору данных [110, 144]. В первом случае экземпляры выборки принадлежат к определенному классу, и эта информация используется в алгоритме дискретизации.

Во втором случае есть только множество значений, которые нужно разделить на интервалы.

Следовательно, для дискретизации случайной величины, характеризующей интенсивность пуассоновского процесса, подходят только алгоритмы дискретизации, работающие с неразмеченным набором данных. Для дискретизации временных интервалов логична дискретизация, отражающая устоявшиеся обозначения временных интервалов (час, день, полдня, неделя, месяц, полгода, год и т.д.). Так как данные об интервалах между эпизодами пуассоновского процесса часто могут быть получены из опросов, есть большая вероятность, что именно такие обозначения будут использовать респонденты. Таким образом для случайных величин $t_{i,j+1}$, $t_{\min}$ и $t_{\max}$ была взята дискретизация вида: $t^{(1)} = [0;0.1)$, $t^{(2)} = [0.1;0.5)$, $t^{(3)} = [0.5;1)$, $t^{(4)} = [1;7)$, $t^{(5)} = [7;30)$, $t^{(6)} = [30;90)$, $t^{(7)} = [90;180)$, $t^{(8)} = [180;\infty)$.

Для дискретизации случайной величины λ были использованы следующие алгоритмы: разбиение на интервалы равные по величине (EW), разбиение на интервалы равные по частоте (EF) и модифицированное разбиение на интервалы равные по частоте (EF_Unique) [144].

EW (Equal width) — это один из самых простых и популярных алгоритмов дискретизации, он делит весь диапазон значений ($R = max - min$) непрерывной величины, на *k* равных интервалов с $k-1$ точками разрыва (c_1, c_2, …, c_{k-1}), которые вычисляются по формуле $c_1 = \min + i \cdot h$, $i = 1,\ldots,k-1$. Длина интервалов (h)— это частное от деления диапазона значений на количество интервалов *k:* $h = R / k$. Недостаток EW заключается в том, что в некоторых интервалах значений может не оказаться, а в некоторых интервалах наоборот и значений будет значительно больше, чем в других.

EF (Equal frequency) — также довольно простой и популярный алгоритм. Он делит диапазон значений на *k* интервалов, каждый из которых содержит около n / k значений, n — это количество всех значений. Алгоритм состоит из следующих

этапов: все значения сортируются в порядке возрастания, делятся на *k* групп, точки разрыва вычисляются как средние арифметическое максимального значения текущей группы значений и минимального значения следующей, далее все непрерывные значения переводятся в значения интервалов, их содержащих. Преимущества этого алгоритма в том, что похожие значения собраны в одном интервале, и уменьшения эффекта резко отличающихся значений, который можно увидеть при применении EW, а то, что два и даже больше близлежащих интервалов могут содержать одинаковые значения, — это его недостаток [110].

EF_Unique — это модифицированный алгоритм EF [144]. EF_Unique состоит из двух этапов: на первом все значения сортируют в восходящем или убывающем порядке, после этого удаляют дублирующие значения, число интервалов *k* устанавливает как ближайшее целое число к квадратному корню из числа оставшихся уникальных значений. На втором этапе значения разделяют на k групп, как в алгоритме EF. Далее высчитывают средние арифметические этих k групп для определения границ интервалов, затем точки разрыва высчитываются как средние арифметические, находящихся друг за другом групп значений. И на последнем шаге непрерывные значения переводят в дискретные с помощью определения интервала, которому они принадлежат.

Еще одной задачей дискретизации непрерывных величин является выбор оптимального числа количества интервалов k. Ведь при маленьком k может быть потеряна некоторая часть информации, а при больших возможно будет довольно трудно правильно интерпретировать полученные результаты [110]. Ниже представлены несколько формул, предложенных для вычисления k [110]: $1+\lfloor \log_2 n \rfloor$ (правило Стерджеса), $\lceil 5\log_{10} n \rceil$ (правило Брукса и Каррузера), $\lceil n^{1/3} \rceil$ (правило Ценкова), $\lceil 2n^{1/3} \rceil$ (правило Райса), $\lceil (2n)^{1/3} \rceil$ (правило Террелла-Скотта).

Учитывая, что изначально генерируется 300 значений интенсивностей (см. раздел 4.4.1), значения k получаются разными при применении разных формул. Рассмотрим 4 значения k: 7, 9, 10, 12 (эти значения могут быть получены при

применении правил Ценкова, Террела–Скотта, Стерджеса, Брукса и Каррузера соответственно).

В EF_Unique число интервалов высчитывается в самом алгоритме, но мы рассмотрим дискретизацию для всех четырех вариантов.

Так как значения интенсивности варьируются от нуля до бесконечности, при дискретизации все интервалы будут начинаться с 0 и заканчиваться бесконечностью.

В таблицах 3.16 – 3.19 представлены результаты дискретизации в виде точек разрыва для всех вариантов k. Кроме применения описанных методов дискретизации также была использована дискретизация на основе экспертных данных (интуитивную, Expert).

Таблица 3.16 — Точки разрыва интервалов дискретизации случайной величины, характеризующй интенсивность пуассоновского процесса при *k=7*

EW	0	0.292	0.583	0.874	1.164	1.455	1.746	∞
EF	0	0.066	0.129	0.202	0.293	0.434	0.673	∞
EF_Unique	0	0.061	0.128	0.207	0.301	0.445	0.782	∞
Expert	0	0.1	0.2	0.3	0.5	0.7	1	∞

Таблица 3.17 — Точки разрыва интервалов дискретизации случайной величины, характеризующй интенсивность пуассоновского процесса при *k=9*

EW	0	0.228	0.454	0.68	0.906	1.132	1.358	1.584	1.81	∞
EF	0	0.049	0.092	0.154	0.212	0.278	0.365	0.514	0.776	∞
EF_Unique	0	0.048	0.097	0.154	0.219	0.292	0.395	0.564	0.912	∞
Expert	0	0.05	0.1	0.2	0.3	0.5	0.7	1	1.5	∞

Таблица 3.18 — Точки разрыва интервалов дискретизации случайной величины, характеризующй интенсивность пуассоновского процесса при *k=10*

EW	0	0.205	0.409	0.612	0.816	1.019	1.222	1.426	1.629	1.833	∞
EF	0	0.045	0.084	0.136	0.189	0.252	0.309	0.417	0.529	0.798	∞
EF_Unique	0	0.043	0.085	0.133	0.189	0.248	0.316	0.415	0.573	0.9	∞
Expert	0	0.05	0.1	0.2	0.3	0.4	0.5	0.7	1	1.5	∞

Таблица 3.19 — Точки разрыва интервалов дискретизации случайной величины, характеризующй интенсивность пуассоновского процесса при *k=12*

EW	0	0.171	0.341	0.51	0.68	0.849	1.019	1.189	1.358	1.528	1.697	1.867	∞
EF	0	0.035	0.074	0.104	0.154	0.192	0.252	0.302	0.365	0.475	0.613	0.867	∞
EF_Unique	0	0.036	0.071	0.108	0.151	0.197	0.247	0.301	0.376	0.479	0.637	0.966	∞
Expert	0	0.01	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1	1.5	∞

Проанализируем работу модели при разных разбиениях случайной величины λ (рассматривается модель оценивания интенсивности пуассоновского процесса, корректно обрабатывающая значение величины интервала между последним эпизодом и окончанием исследования). На полученных данных было проведено обучение параметров БСД. То есть для каждой пары переменных, соединенных ребром, были вычислены условные вероятности. Таким образом было получено 12 моделей, каждая обученная при различном разбиении случайной непрерывной величины λ, характеризующей интенсивность поведения.

Основная задача построенной модели — автоматизация получения оценки интенсивности пуассоновского процесса при доступности сведений только о длине интервалов между несколькими последними эпизодами, а также минимальном и максимальном значении длин интервалов между эпизодами. При дискретизации значений интенсивности — это задача классификации по k классам. Таким образом, одна из главных характеристик работы модели в данном случае является средняя точность. Рассмотрим точность (acc.) и среднюю точность модели при различных разбиениях величины λ. В таблице 3.20 показаны полученные результаты.

Таблица 3.20 — Результаты работы модели при различных дискретизациях λ [59]

Метод дискретизации	**Количество интервалов**	**Точность**	**Средняя точность**
EW	7	0.729	0.923
EW	9	0.648	0.922
EW	10	0.627	0.925

EW	12	0.587	0.931
EF	7	0.48	0.851
EF	9	0.437	0.875
EF	10	0.385	0.877
EF	12	0.347	0.891
EF_Unique	7	0.507	0.859
EF_Unique	9	0.414	0.87
EF_Unique	10	0.392	0.878
EF_Unique	12	0.31	0.885
Expert	7	0.535	0.867
Expert	9	0.488	0.886
Expert	10	0.447	0.889
Expert	12	0.424	0.904

Выводы по результатам вычислительного эксперимента: самые высокие результаты показала модель, обученная при дискретизации алгоритмом EW при $k = 7$. Самая низкая точность была получена при EF_Unique и $k = 12$ (0.31), а самая низкая средняя точность при EF и $k = 7$ (0.851). Дискретизация с помощью алгоритма EW в целом показала лучшие результаты, однако надо учитывать, что при использовании EW и работе модели на реальных данных можно получить ухудшение этих показателей из-за большего числа сильно отличающихся экземпляров процесса. Экспертная дискретизация также показала довольно хорошие результаты, что значит, что при использовании моделей оценивания интенсивности пуассоновского процесса в различных областях науки на этом этапе исследований можно обратиться к эксперту определенной области. Применение EF_Unique не дало значительных положительных изменений по сравнению с EF. Заметим, что при увеличении количества интервалов уменьшается точность, но средняя точность наоборот увеличивается, а в данном исследовании именно она считается основной метрикой эффективности модели. Можно отметить, что эта метрика является довольно высокой во всех случаях и изменяется в промежутке от 0.851 до 0.923, что говорит о возможности применять данную модель при

различных дискретизациях величины, характеризующей интенсивность пуассоновского процесса.

ВЫВОДЫ ПО ГЛАВЕ 3

В третьей главе представлены основные теоретические результаты данного диссертационного исследования. В первом разделе рассмотрены метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанной на БСД. Во втором представлен алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, рассмотрены модели, расширенные скрытыми переменными, соответствующими истинным сведениям об эпизодах и рекордных интервалах пуассоновского процесса. В третьем разделе представлены метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений. В четвертом разделе представлены модели оценивания интенсивности постинга в социальной сети, расширенные за счет объективных данных о пользователе. В пятом разделе рассмотрены возможные варианты дискретизации непрерывных величин, входящих в модели оценивания интенсивности пуассоновского процесса.

В предложенных методах и алгоритмах учитываются следующие виды неопределенности, возникающие при оценивании интенсивности пуассоновского процесса: 1. Несогласованность ответов респондентов (первый раздел). 2. Некорректность ответов респондентов (второй раздел). 3. Некорректность задания момента окончания исследования (третий раздел).

ГЛАВА 4. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И АПРОБАЦИЯ

В третьей главе были представлены методы и алгоритмы для получения оценки интенсивности пуассоновского процессов исходя из данных о последних эпизодах и рекордных интервалах пуассоновского процесса, обрабатывающие возникающие при этом различные виды неопределенности. В этой главе описаны архитектура и прототип комплекса программ для работы с моделями оценивания интенсивности пуассоновского процесса, который обеспечивает работу с предложенными методами и алгоритмами для специалистов из различных областей, изучающих поведение человека, реализующего работу разработанных методов и алгоритмов.

В прототип комплекса программ для работы с моделями оценивания интенсивности пуассоновского процесса входят три модуля:

- модуль для работы с инструментом оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса;
- модуль для работы с моделью оценивания интенсивности пуассоновского процесса со скрытыми переменными;
- модуль для работы с моделью оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом.

Архитектура прототипа комплекса программ для работы с моделями оценивания интенсивности пуассоновского процесса, который обеспечивает работу с предложенными методами и алгоритмами для специалистов из различных областей, изучающих поведение человека изображена на рис. 4.1.

Предложенный прототип комплекса может применяться для работы с моделями оценивания интенсивности пуассоновского процесса, проведения экспериментов и для получения оценки интенсивности пуассоновского процессов поведения в различных исследованиях.

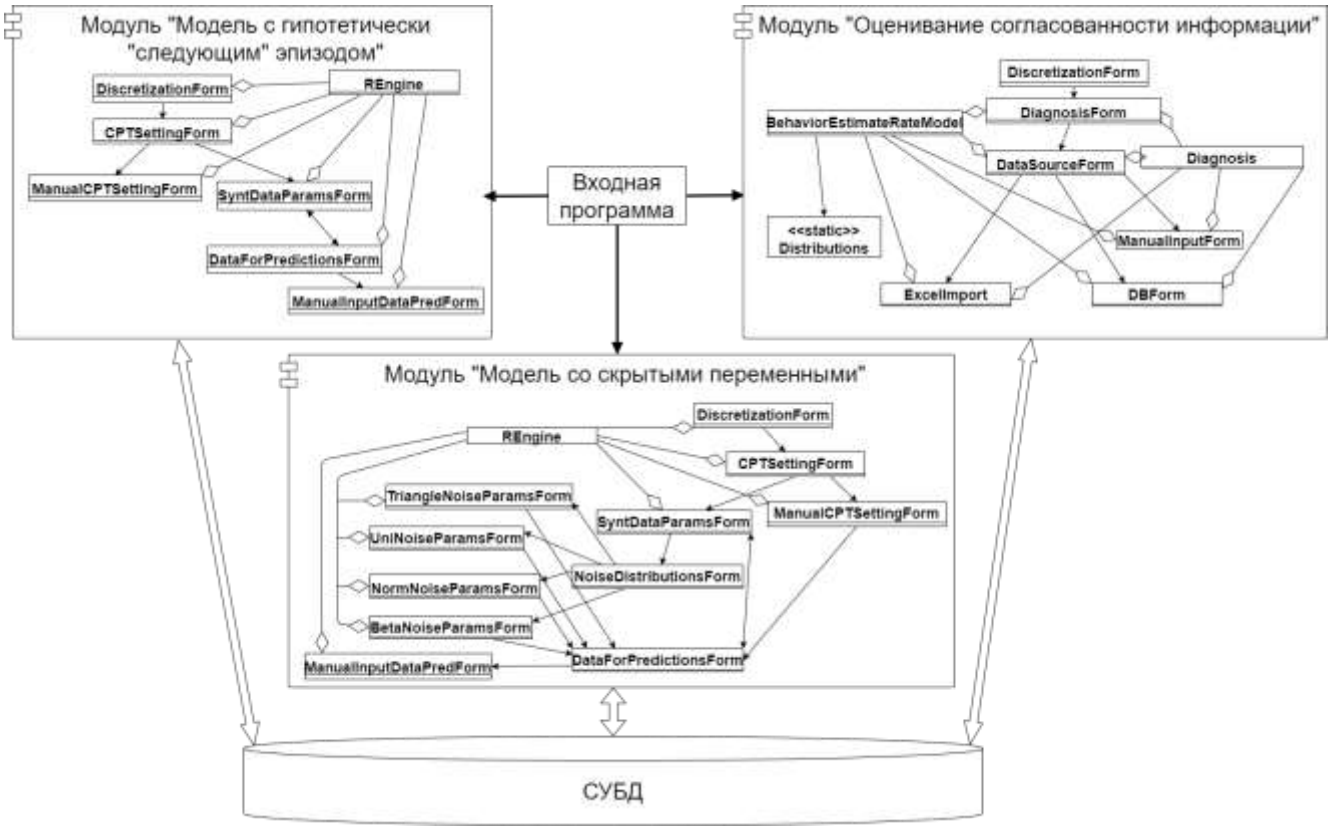

Рисунок 4.1 — Архитектура прототипа комплекса программ для работы с моделями оценивания интенсивности пуассоновского процесса

Кроме того, описаны собранные данные (синтетические и данные из социальных сетей) для апробации предложенных моделей, результаты апробации, а также описывается их применение.

4.1 МОДУЛЬ ДЛЯ РАБОТЫ С ИНСТРУМЕНТОМ ОЦЕНИВАНИЯ СОГЛАСОВАННОСТИ ИНФОРМАЦИИ О ПОСЛЕДНИХ ЭПИЗОДАХ И РЕКОРДНЫХ ИНТЕРВАЛАХ ПУАССОНОВСКОГО ПРОЦЕССА

Для более удобной работы с предложенным методом диагностики согласованности было разработано программное обеспечение. При разработке использовались C# и библиотека Smile [92] , СУБД MySQL. Использовалась среда разработки Microsoft Visual Studio.

На рисунке 4.2 представлена диаграмма классов модуля для работы с инструментом оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса.

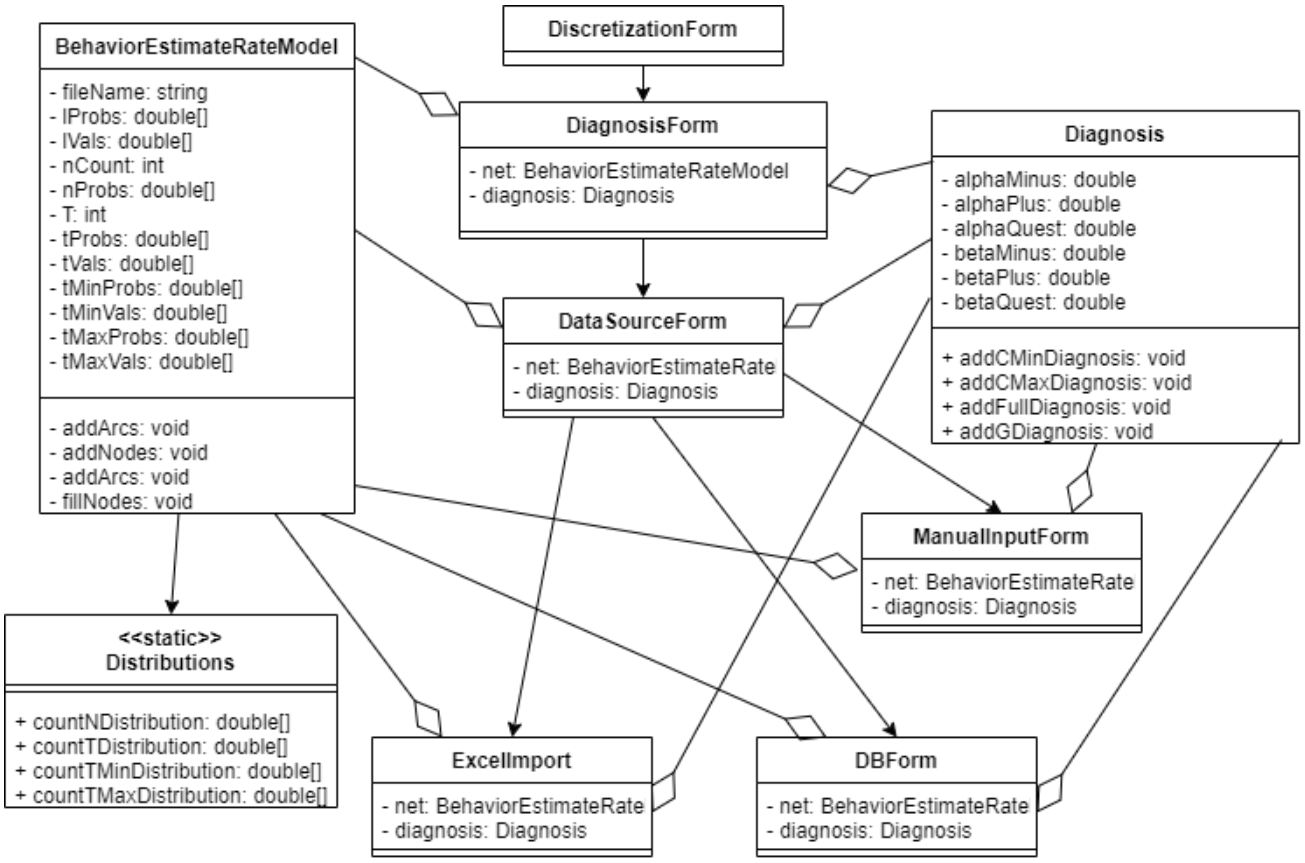

Рисунок 4.2 — Диаграмма классов модуля оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса

Работа с модулем представляет собой последовательное заполнение форм: DiscretizationForm, в которой нужно задать интервалы дискретизации непрерывных величин; DiagnosisForm, при работе с которой в построенную модель оценивания интенсивности пуассоновского процесса (BehaviorEstimateRateModel) добавляются вершины, характеризующие согласованность данных респондентов (Diagnosis); DataSourceForm, в которой нужно выбрать данные, по которым будут строиться оценки интенсивности пуассоновского процесса, можно ввести данные вручную (ManualInputForm), выбрать импорт данных из файла Excel (ExcelImport) или базы данных (DBForm), в первом случае оценки интенсивности появятся прямо в форме, во втором сохранятся в файл Excel, в третьем сохранятся в базе данных.

На рисунках 4.3 – 4.4 представлены фрагменты интерфейса. В начале работы пользователю предлагается определить интервалы t_{ij}, $t_{\min}$ и $t_{\max}$ (рис. 4.3).

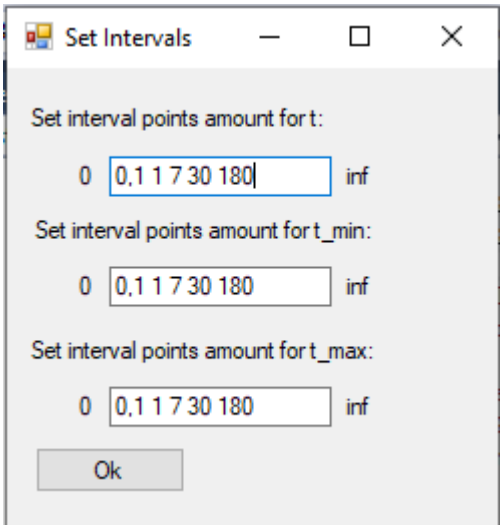

Рисунок 4.3 — Определение дискретизации непрерывных величин [39]

После того, как заданы эти интервалы, можно задать значения параметрам $\alpha^{(s)}$ и $\beta^{(s)}$, $s \in \{+,-,?\}$, а затем дополнить модель инструментом оценивания согласованности данных, что можно сделать, постепенно добавляя вершины, характеризующие согласованность данных, или сразу (рис. 4.3).

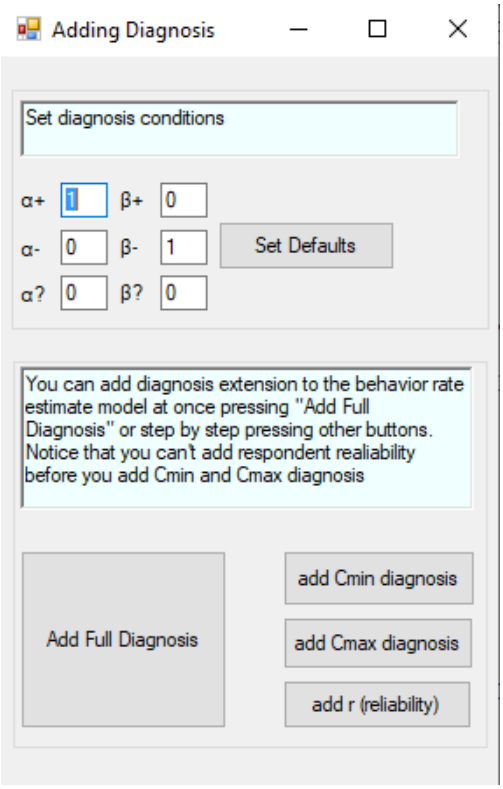

Рисунок 4.4 — Добавление инструмента оценивания согласованности ответов респондентов к модели [39]

На следующем этапе можно ввести данные, полученные от респондентов. Есть несколько вариантов ввода (рис. 4.5). Это можно сделать вручную (рис. 4.6), с помощью электронной таблицы MS Excel или базы данных MySQL. При выборе ввода данных из файла результаты диагностики также сохраняются в отдельном файле.

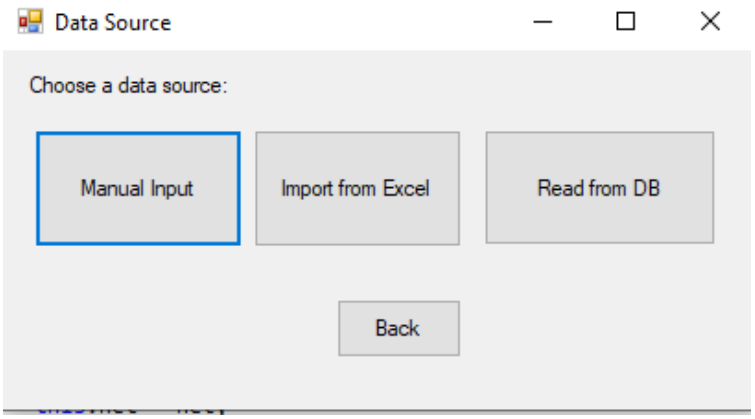

Рисунок 4.5 — Выбор источника данных

При вводе данных вручную пользователь указывает интервалы (выбор из выпадающего списка) для последних трех эпизодов поведения, а также максимальный и минимальный интервалы между эпизодами. После нажатия кнопки «Get Diagnosis» результат диагностики согласованности вершин появляется в специально предусмотренном для этого поле (рис. 4.6).

На программу было получено свидетельство о государственной регистрации программы для ЭВМ [62].

Рассмотрим пример работы с программой. Разобьем t_{ij}, $t_{\min}$ и $t_{\max}$ на интервалы, предложенные по умолчанию (см. рис. 4.3), то есть $(0;0.1)$, $[0.1;1)$, $[1;7)$, $[7;30)$, $[30;180)$ и $[180;+\infty)$. При заданных по умолчанию параметрах, то есть $\alpha^{+}=1,\ \alpha^{-}=0,\ \alpha^{?}=0$ и $\beta^{+}=0,\ \beta^{-}=1,\ \beta^{?}=0$, вероятность согласованности данных респондента будет равна нулю, если данные, противоречат друг другу, и единице, при отсутствии противоречий.

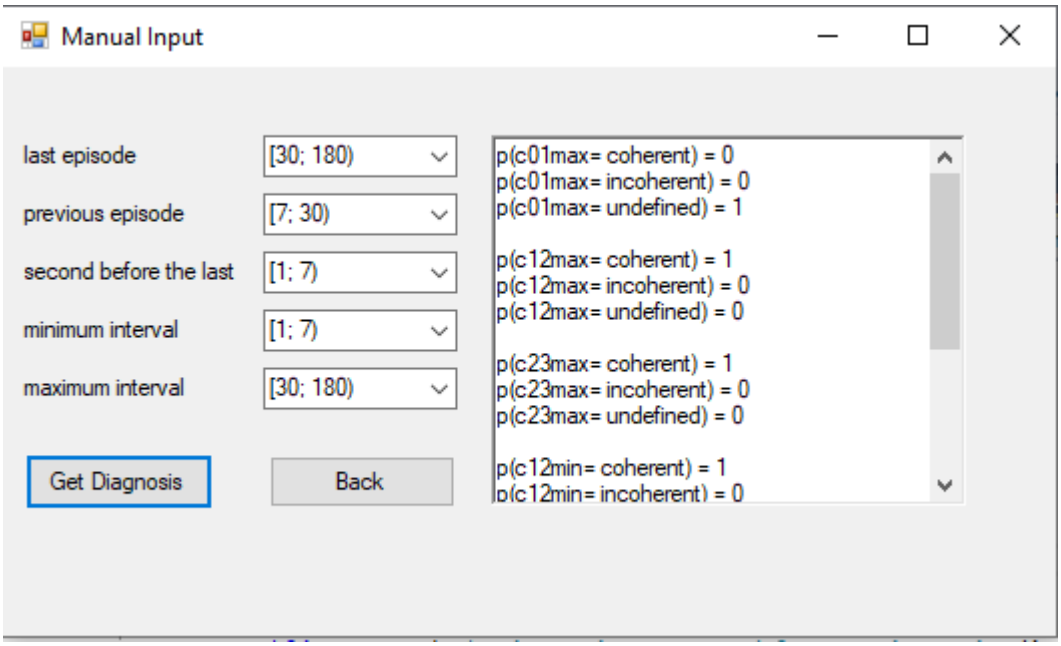

Рисунок 4.6 — Диагностика согласованности ответов респондента при ручном вводе данных [39]

Данные респондентов представлены в таблице 4.1.

Таблица 4.1 — Данные респондентов

респондент	t_{01}	t_{12}	t_{23}	$t_{\min}$	$t_{\max}$
1	5	5	5	5	5
2	3	0.2	4	0.1	10
3	0.5	4	0.5	0.05	20
4	10	0.5	10	5	60
5	7	1	0.05	0.05	7

В таблице 4.2 показаны результаты диагностики.

Таблица 4.2 — Результаты диагностики I

№	$c^{+}_{t_{01,\max}}$	$c^{-}_{t_{01,\max}}$	$c^{?}_{t_{01,\max}}$	$c^{+}_{t_{12,\max}}$	$c^{-}_{t_{12,\max}}$	$c^{?}_{t_{12,\max}}$	$c^{+}_{t_{23,\max}}$	$c^{-}_{t_{23,\max}}$	$c^{?}_{t_{23,\max}}$
1	0	0	1	0	0	1	0	0	1
2	1	0	0	1	0	0	1	0	0
3	1	0	0	1	0	0	1	0	0
4	1	0	0	1	0	0	1	0	0
5	0	0	1	1	0	0	1	0	0
№	$c^{+}_{t_{12,\min}}$	$c^{-}_{t_{12,\min}}$	$c^{?}_{t_{12,\min}}$	$c^{+}_{t_{23,\min}}$	$c^{-}_{t_{23,\min}}$	$c^{?}_{t_{23,\min}}$	r^{+}	r^{-}	$r^{?}$

1	0	0	1	0	0	1	0	0	1
2	0	0	1	1	0	0	0.8	0	0.2
3	1	0	0	1	0	0	1	0	0
4	0	1	0	1	0	0	0.8	0.2	0
5	1	0	0	0	0	1	0.6	0	0.4

Теперь изменим некоторые параметры, пусть $\alpha^{+}=0.9$, $\alpha^{-}=0.1$, $\alpha^{?}=0$, $\beta^{+}=0.1$, $\beta^{-}=0.9$ и $\beta^{?}=0$. Результаты диагностики с такими параметрами можно увидеть в таблице 4.3.

Таблица 4.3 — Результаты диагностики II

респондент	$c^{+}_{t_{01,\max}}$	$c^{-}_{t_{01,\max}}$	$c^{?}_{t_{01,\max}}$	$c^{+}_{t_{12,\max}}$	$c^{-}_{t_{12,\max}}$	$c^{?}_{t_{12,\max}}$	$c^{+}_{t_{23,\max}}$	$c^{-}_{t_{23,\max}}$	$c^{?}_{t_{23,\max}}$
1	0	0	1	0	0	1	0	0	1
2	0.9	0.1	0	0.9	0.1	0	0.9	0.1	0
3	0.9	0.1	0	0.9	0.1	0	0.9	0.1	0
4	0.9	0.1	0	0.9	0.1	0	0.9	0.1	0
5	0	0	1	0.9	0.1	0	0.9	0.1	0
респондент	$c^{+}_{t_{12,\min}}$	$c^{-}_{t_{12,\min}}$	$c^{?}_{t_{12,\min}}$	$c^{+}_{t_{23,\min}}$	$c^{-}_{t_{23,\min}}$	$c^{?}_{t_{23,\min}}$	r^{+}	r^{-}	$r^{?}$
1	0	0	1	0	0	1	0	0	1
2	0	0	1	0.9	0.1	0	0.656	0.144	0.2
3	0.9	0.1	0	0.9	0.1	0	0.82	0.18	0
4	0.1	0.9	0	0.9	0.1	0	0.692	0.308	0
5	0.9	0.1	0	0	0	1	0.492	0.108	0.4

С полученными результатами можно действовать по-разному в зависимости от условий конкретной задачи, например, исключить из рассмотрения данные, не удовлетворяющие определенному порогу согласованности.

4.2 МОДУЛЬ ДЛЯ РАБОТЫ С МОДЕЛЬЮ ОЦЕНИВАНИЯ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА СО СКРЫТЫМИ ПЕРЕМЕННЫМИ

Для работы с моделью оценивания интенсивности пуассоновского процесса со скрытыми переменными было разработано программное обеспечение. При разработке использовались C#, R [198] и пакет bnlearn [98], библиотека RDotNet [201] обеспечила взаимодействие между двумя языками. Использовались среды разработки Microsoft Visual Studio и RStudio [209].

На рис. 4.7 представлена диаграмма классов модуля для работы с моделью оценивания интенсивности пуассоновского процесса со скрытыми переменными. Благодаря классу REngine происходит взаимодействие с языком R, все этапы работы с моделью оценивания интенсивности пуассоновского процесса написаны на языке R с использованием пакета bnlearn. Работа с модулем представляет собой последовательное заполнение форм: DiscretizationForm, в которой нужно задать интервалы дискретизации непрерывных величин; CPTSettingForm, в которой надо выбрать вариант ввода данных для построения тензоров условной вероятности модели, можно выбрать заполнение таблиц условной вероятности вручную (ManualCPTSettingForm), можно выбрать статистические данные для обучения модели из файла или синтезировать данные внутри программы (данные синтезированные программой сохраняются в файлы формата csv или Excel), при синтезе данных надо указать нужные параметры в SyntDataParamForm, так как для обучения модели используются также зашумленные данные, характеризующие ответы респондентов, также надо выбрать параметры зашумления данных, а именно выбрать тип распределения для генерации шума (NoiseDistributionsForm) и задать необходимые параметры, для этого есть 4 варианта: треугольное распределение (TriangleNoiseParamsForm), равномерное (UniNoiseParamsForm), нормальное (NormNoiseParamsForm) и бета-распределение (BetaNoiseParamsForm). После обучения модели можно приступать к получению оценок интенсивности пуассоновского процесса по данным, DataForPredictionsForm предлагает выбрать

данные из файла (csv или Excel) или ввести данные вручную (ManualInputDataPredForm). в первом случае оценки интенсивности сохранятся в файл, определенный пользователем, во втором —появятся прямо в форме.

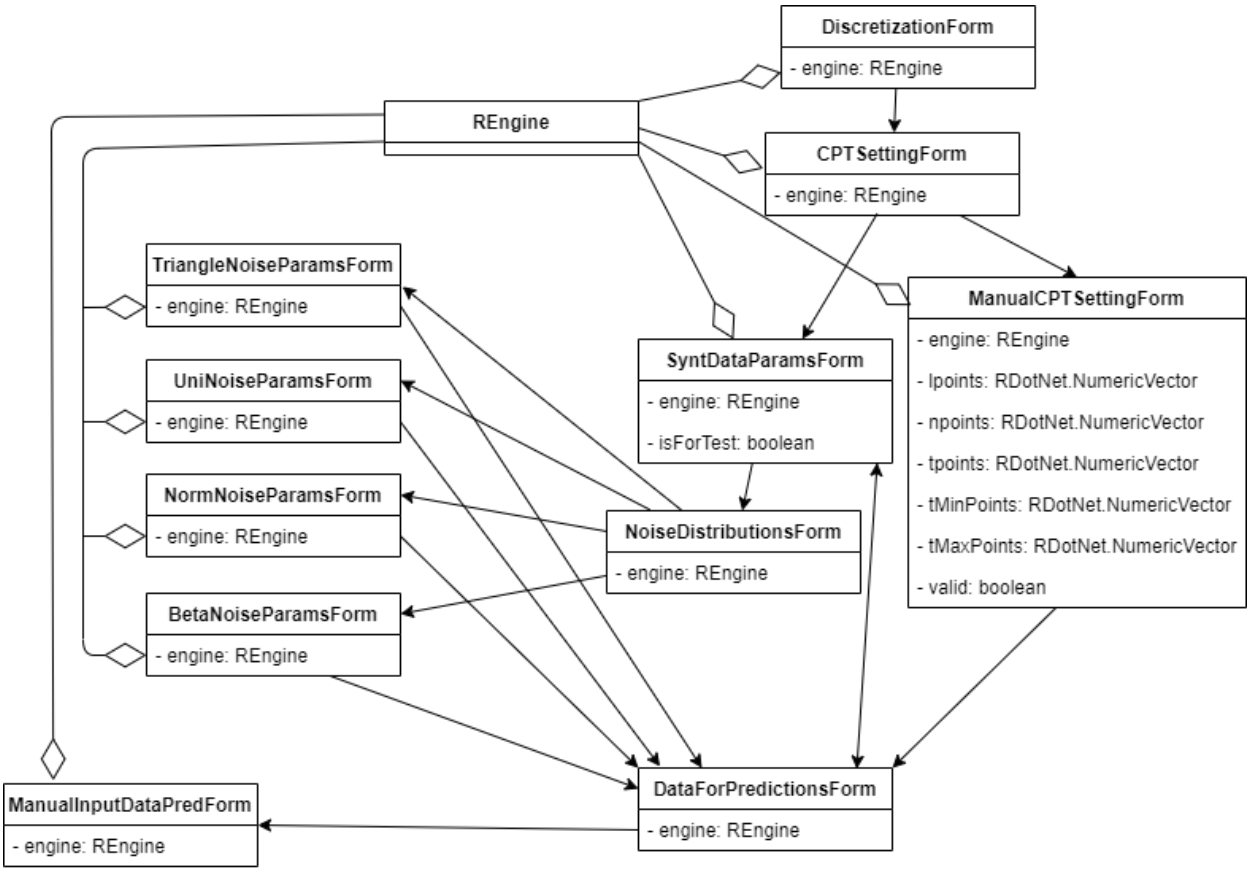

Рисунок 4.7 — Диаграмма классов модуля для работы с моделью оценивания интенсивности пуассоновского процесса со скрытыми переменными

Вначале надо определить интервалы для дискретизации для всех непрерывных величин, то есть для величин, обозначающих временные интервалы и интенсивность, для этого надо перечислить точки разрыва, по умолчанию эти поля заполнены, но можно установить свои значения (рис. 4.8).

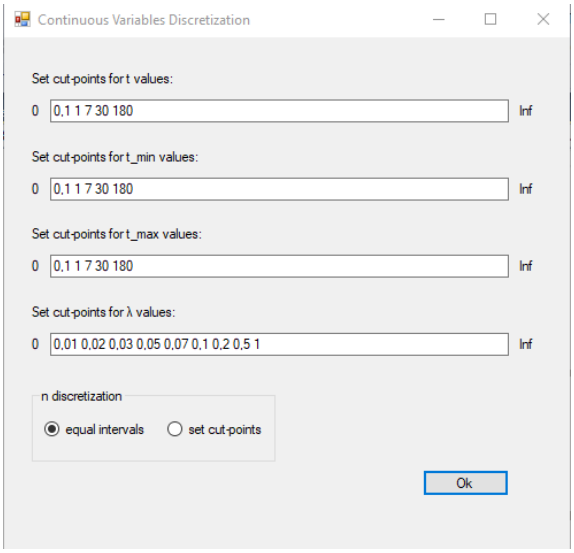

Рисунок 4.8 — Определение дискретизации непрерывных величин

Также надо выбрать вид дискретизации величины n, характеризующей количество эпизодов, произошедших в течение исследуемого периода, для нее можно также установить точки разрыва или выбрать дискретизацию на равные промежутки, указав их количество (рис. 4.9).

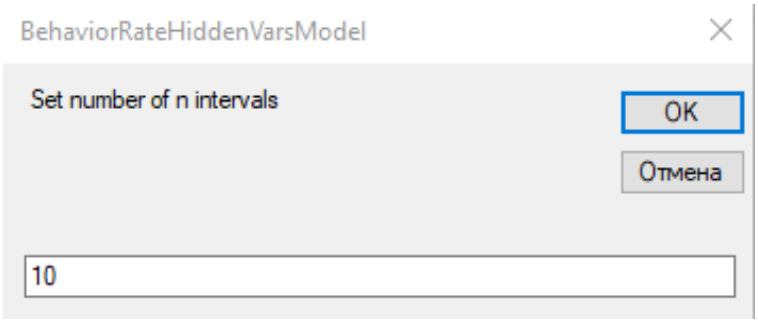

Рисунок 4.9 — Определение числа интервалов для дискретизации n

После этого надо обучить модель, то есть установить априорные вероятности, это можно сделать тремя способами (рис. 4.10): используя синтетические данные, сгенерированные программой, вручную заполнив таблицы условных вероятностей (рис. 4.18), или выбрав файл с данными для обучения в формате таблиц Excel или csv.

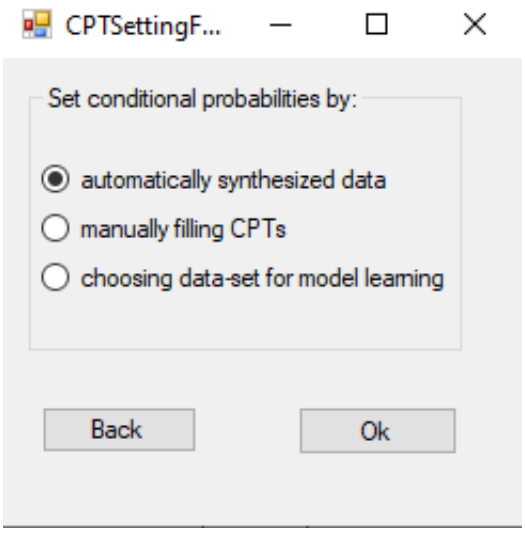

Рисунок 4.10 — Выбор данных для обучения модели оценки интенсивности пуассоновского процесса со скрытыми переменными

При выборе синтетических данных появится окно с параметрами, нужными для синтеза данных (рис. 4.11).

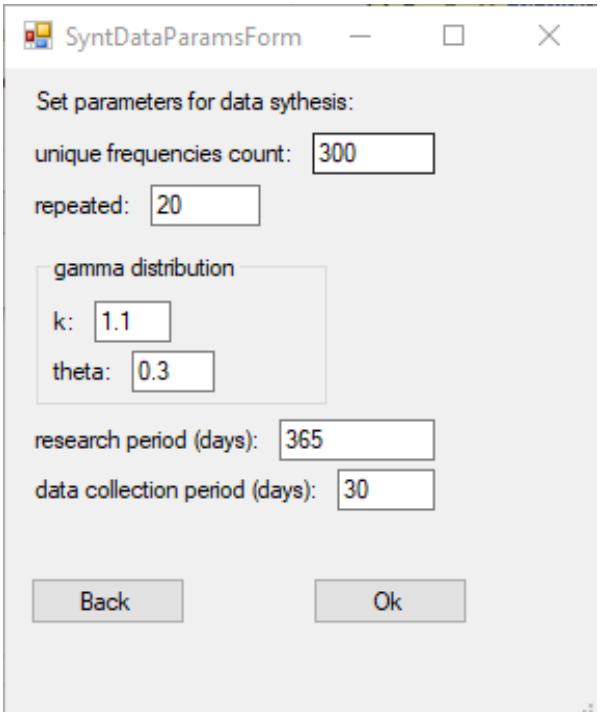

Рисунок 4.11— Установка параметров синтеза данных для обучения модели оценки интенсивности пуассоновского процесса со скрытыми переменными

После определения этих параметров нужно выбрать распределения для синтеза «зашумленных» ответов респондентов. Для каждой переменной, соответствующей ответу респондента надо выбрать одно из четырех

распределений (рис. 4.12) и установить соответствующие параметры: для нормального — математическое ожидание и стандартное отклонение (рис. 4.15), если полученное синтезированное значение окажется меньше нуля, оно будет заменено на 0), для треугольного — нижнюю и верхнюю границы (рис. 4.13), для бета-распределения — параметры формы (рис. 4.16), для равномерного распределения надо указать отклонение, то есть на сколько «зашумленное» значение будет отличаться от исходного (рис. 4.14).

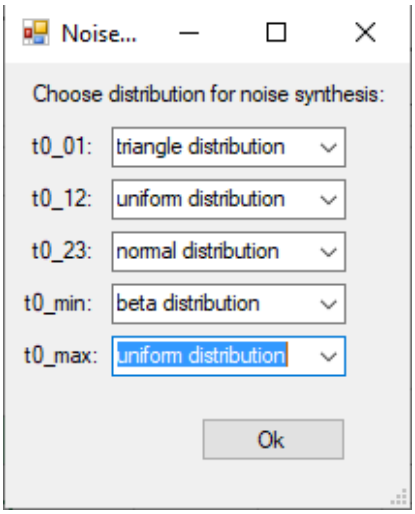

Рисунок 4.12 — Выбор распределения для синтеза «зашумленных» ответов респондентов

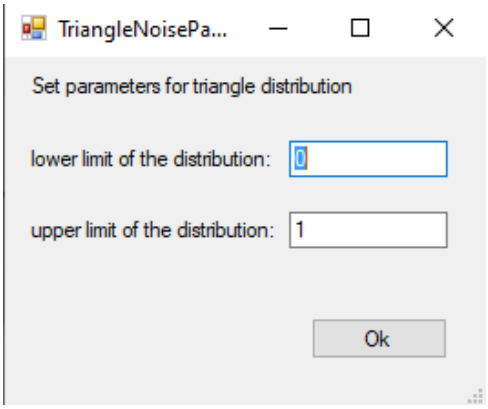

Рисунок 4.13 — Установка параметров для треугольного распределения для синтеза «зашумленных» ответов респондентов

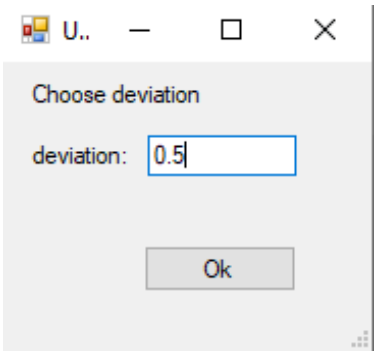

Рисунок 4.14 — Установка отклонения от ответов респондента при выборе равномерного распределения для синтеза «зашумленных» ответов респондентов

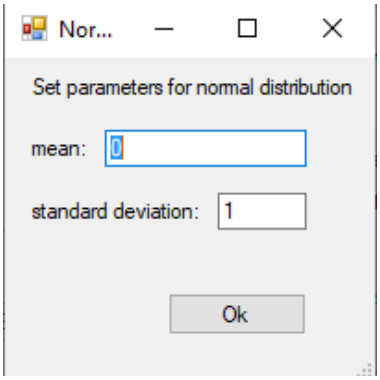

Рисунок 4.15 — Установка параметров для нормального распределения для синтеза «зашумленных» ответов респондентов

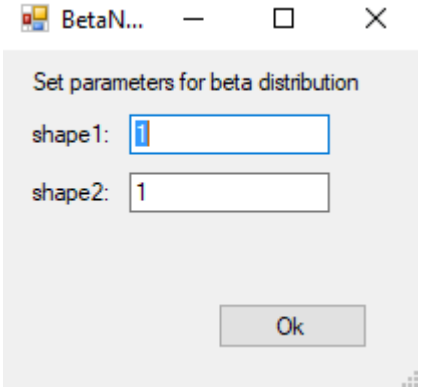

Рисунок 4.16 — Установка параметров для бета-распределения для синтеза «зашумленных» ответов респондентов

После установки всех параметров (рис. 4.17) модель со скрытыми переменными обучается на полученном синтезированном множестве данных, также этот набор данных сохраняется в отдельном файле.

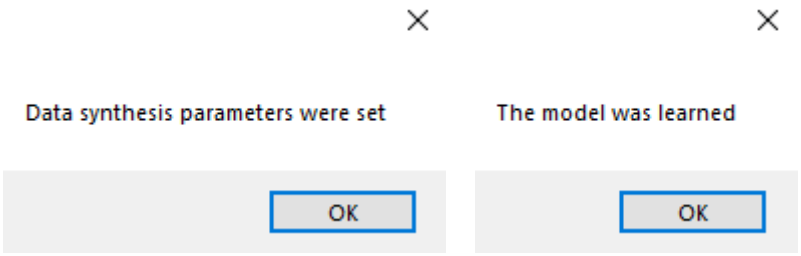

Рисунок 4.17 — Всплывающие сообщения о ходе процесса работы с моделью оценки интенсивности пуассоновского процесса со скрытыми переменными

Если пользователь выбрал заполнение таблиц условных вероятностей вручную, появляется окно со вкладками для каждой вершины модели (рис. 4.18), внутри каждой вкладки находится таблица для заполнения.

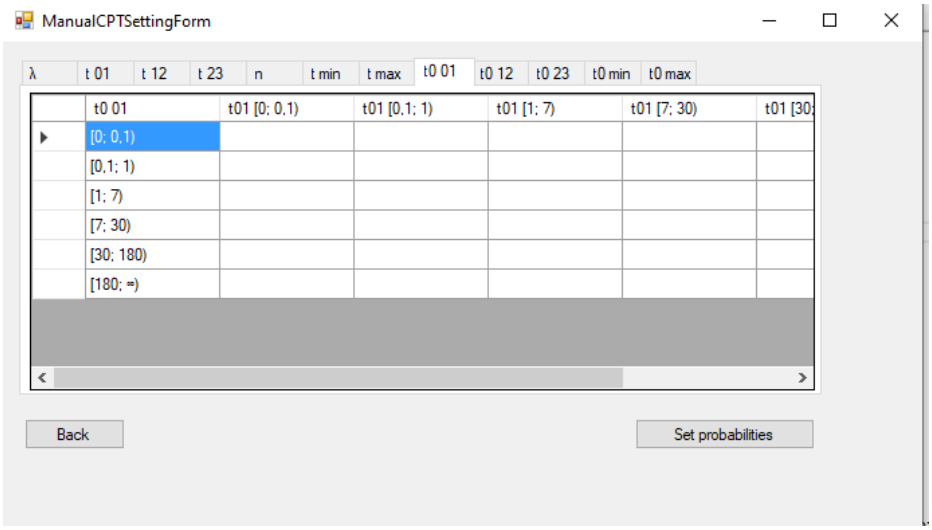

Рисунок 4.18 — Заполнение таблиц условных вероятностей вручную

После заполнении таблиц условных вероятностей производится проверка их заполненности, а также того, что сумма вероятностей для каждого значения должна быть равной 1. На рисунке 4.19 приведен пример предупреждающего сообщения.

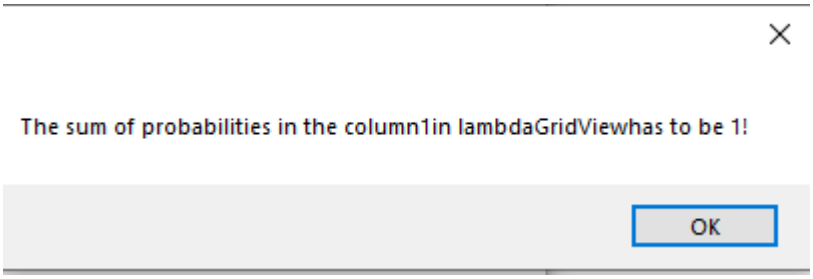

Рисунок 4.19 — Предупреждающее сообщение об ошибке при заполнении таблиц условных вероятностей

После того как модель была обучена, то есть заполнены все таблицы условных вероятностей, можно приступать к оцениванию интенсивности пуассоновского процесса. Данные по которым нужно получить оценку также можно ввести тремя способами (рис. 4.20): можно синтезировать набор данных для того, чтобы проверить работу модели (при этом будут известны исходные интенсивности), можно ввести данные вручную, также можно выбрать файл формата csv или электронных таблиц Excel.

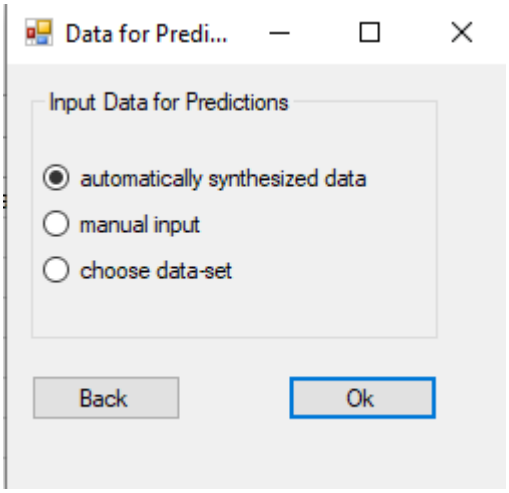

Рисунок 4.20 — Выбор данных для модели оценки интенсивности пуассоновского процесса со скрытыми переменными

При выборе ввода данных вручную пользователь должен ввести данные об интервалах между эпизодами поведения в днях, по нажатию кнопки «Predict Behavior Rate» в соответствующем поле появляется оценка интенсивности поведения с заданной пользователем дискретизацией (рис. 4.21).

При выборе синтетических данных пользователь снова должен указать все необходимые параметры (рис.4.11–4.16), после этого результаты работы модели сохраняются в файл формата csv или таблиц Excel в указанном пользователем месте.

При выборе файла пользователь может выбрать файл с данными формата таблиц Excel или csv с помощью окна выбора файлов, после того как модель получит оценки для всех записей, файл с результатами также сохраняется в формате csv или таблиц Excel в указанном пользователем месте.

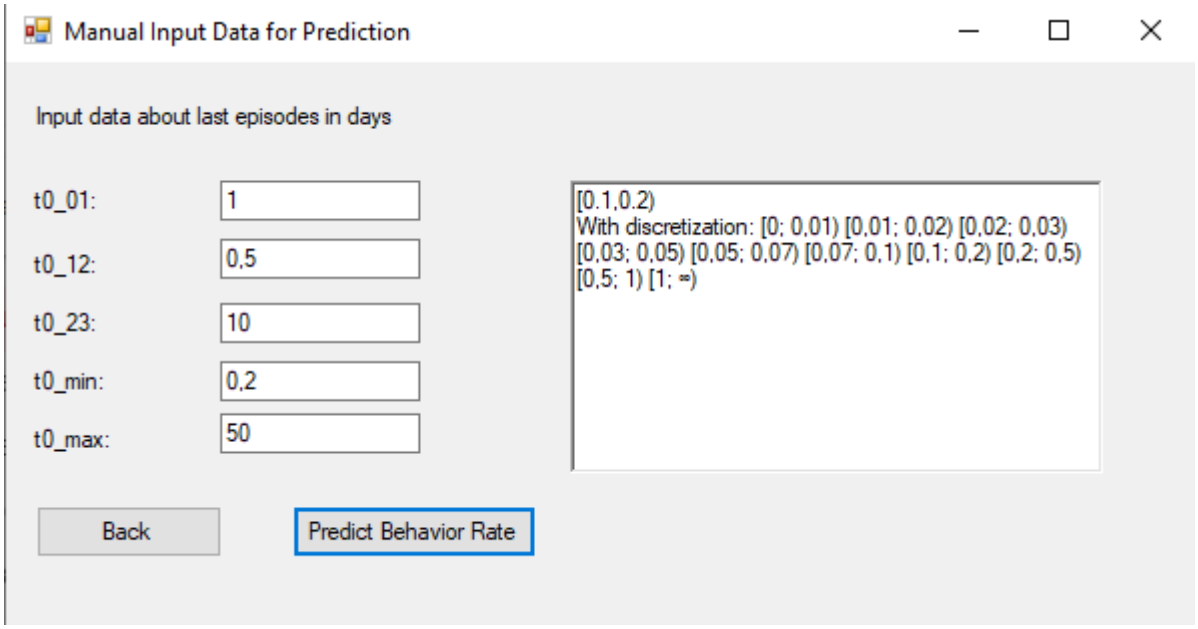

Рисунок 4.21 — Оценивание интенсивности пуассоновского процесса моделью при ручном вводе данных

4.3 МОДУЛЬ ДЛЯ РАБОТЫ С МОДЕЛЬЮ ОЦЕНИВАНИЯ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА С ГИПОТЕТИЧЕСКИ «СЛЕДУЮЩИМ» ЭПИЗОДОМ

Для работы с моделью оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом, на которой основаны метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений, также была написана программа. Программа написана на двух языках: часть, отвечающая за взаимодействие с пользователем реализована на языке C#, а часть, отвечающая за работу с моделью, то есть работу с БСД, реализована на языке R [198] с использованием пакета bnlearn [98]. Взаимодействие между этими частями обеспечено с помощью библиотеки RDotNet [201]. Использовались среды разработки Microsoft Visual Studio и RStudio [209].

На рис. 4.22 представлена диаграмма классов модуля для работы с моделью оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом. На рис. 4.7 представлена диаграмма классов модуля для

работы с моделью оценивания интенсивности пуассоновского процесса со скрытыми переменными. Благодаря классу REngine происходит взаимодействие с языком R, все этапы работы с моделью оценивания интенсивности пуассоновского процесса написаны на языке R с использованием пакета bnlearn. Работа с модулем представляет собой последовательное заполнение форм: DiscretizationForm, в которой нужно задать интервалы дискретизации непрерывных величин; CPTSettingForm, в которой надо выбрать вариант ввода данных для построения тензоров условной вероятности модели, можно выбрать заполнение таблиц условной вероятности вручную (ManualCPTSettingForm), можно выбрать статистические данные для обучения модели из файла или синтезировать данные внутри программы (данные синтезированные программой сохраняются в файлы формата csv или Excel), при синтезе данных надо указать нужные параметры в SyntDataParamForm. После обучения модели можно приступать к получению оценок интенсивности пуассоновского процесса по данным, DataForPredictionsForm предлагает выбрать данные из файла (csv или Excel) или ввести данные вручную (ManualInputDataPredForm). в первом случае оценки интенсивности сохранятся в файл, определенный пользователем, во втором — появятся прямо в форме.

На первом этапе надо определить интервалы для дискретизации для всех непрерывных величин, то есть для величин, обозначающих временные интервалы и интенсивность, для этого надо перечислить точки разрыва, по умолчанию эти поля заполнены, но можно установить свои значения (рис. 4.23).

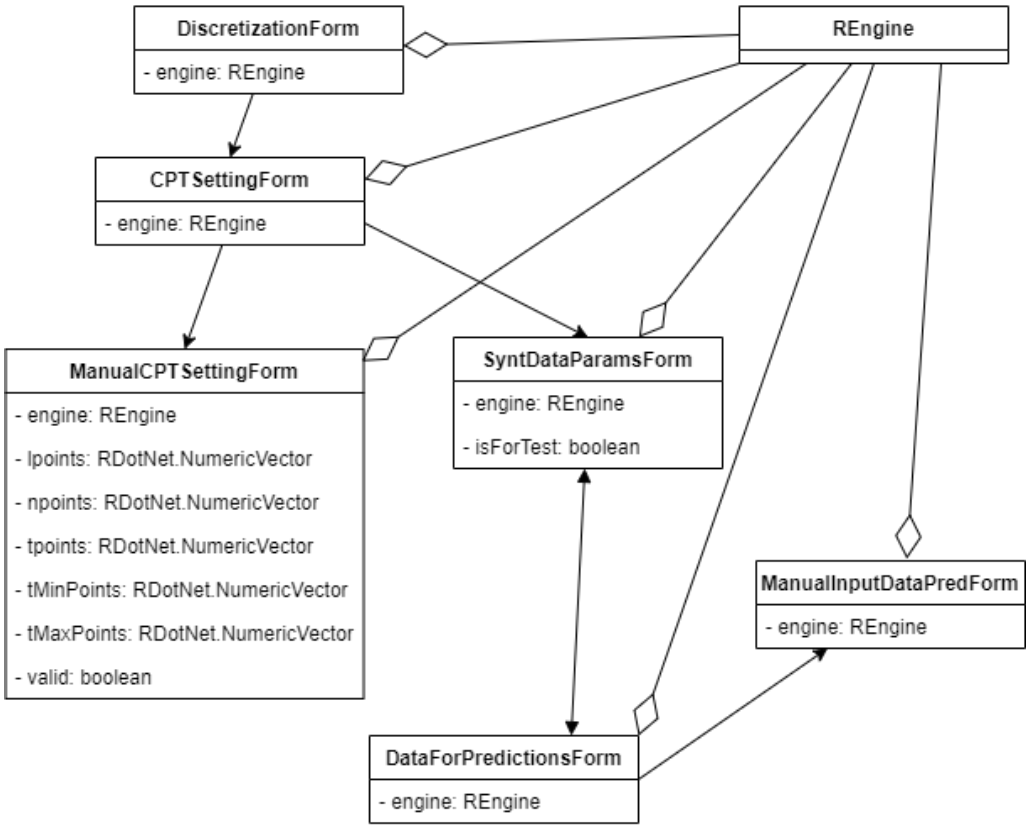

Рисунок 4.22 — Диаграмма классов модуля для работы с моделью оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом

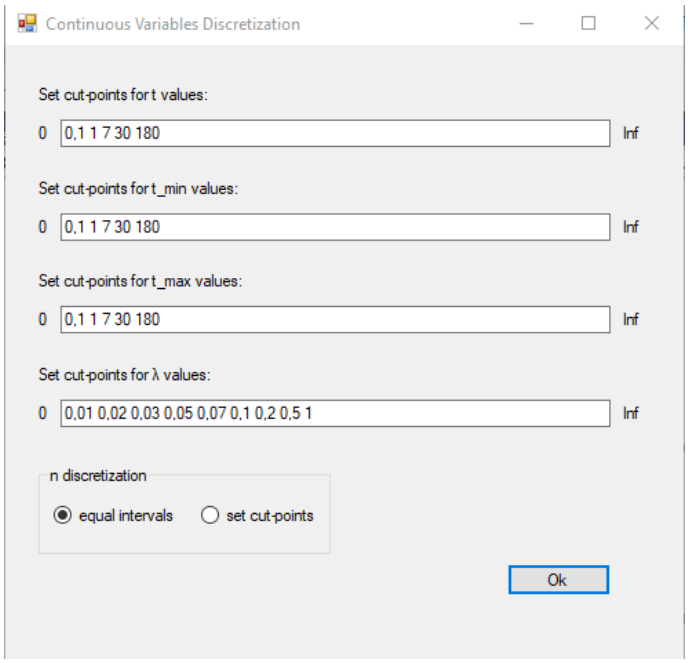

Рисунок 4.23 — Определение дискретизации непрерывных величин

Также надо выбрать вид дискретизации величины *n*, характеризующей количество эпизодов, произошедших в течение исследуемого периода, для нее можно также установить точки разрыва или выбрать дискретизацию на равные промежутки, указав их количество в появляющемся при соответствующем выборе окне диалога.

После этого надо обучить модель, то есть установить априорные вероятности, это можно сделать тремя способами (рис. 4.24): используя синтетические данные, сгенерированные программой, вручную заполнив таблицы условных вероятностей (рис. 4.27), или выбрав файл с данными для обучения в формате таблиц Excel или csv.

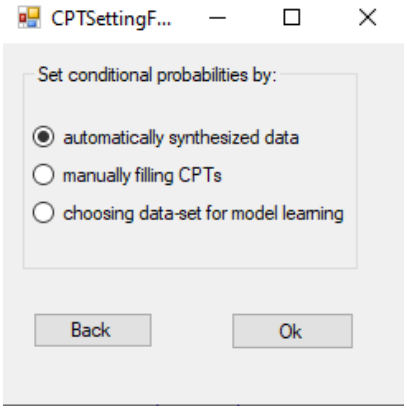

Рисунок 4.24 — Выбор данных для обучения модели оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом

При выборе синтетических данных появится окно с параметрами, нужными для синтеза данных (рис. 4.25).

После установки всех параметров (рис. 4.26) модель обучается на полученном синтезированном множестве данных, также этот набор данных сохраняется в отдельном файле.

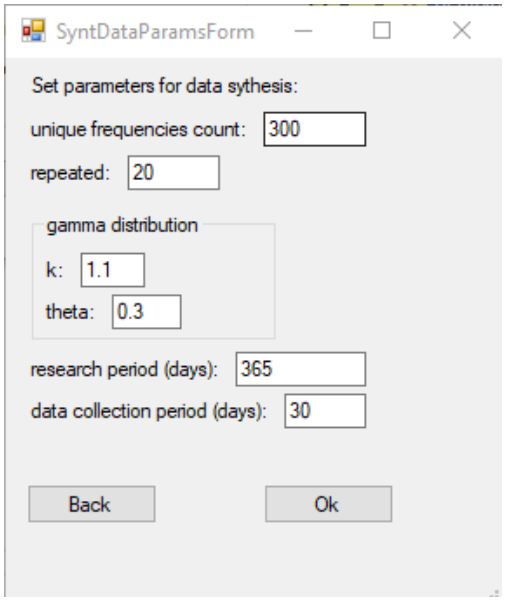

Рисунок 4.25 — Установка параметров синтеза данных для обучения модели оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом

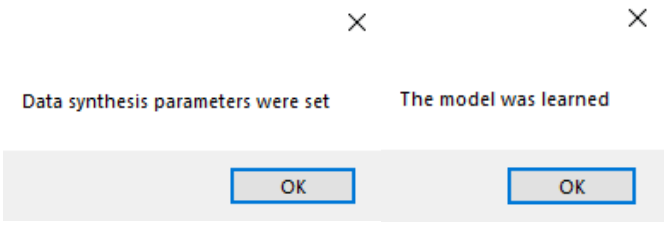

Рисунок 4.26 — Всплывающие сообщения о ходе процесса работы с моделью оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом

Если пользователь выбрал заполнение таблиц условных вероятностей вручную, появляется окно со вкладками для каждой вершины модели (рис. 4.27), внутри каждой вкладки находится таблица для заполнения.

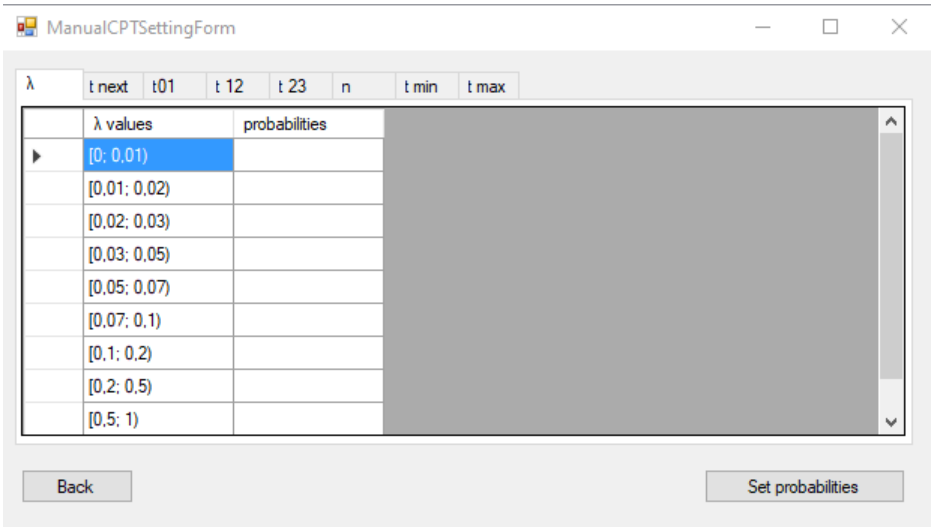

Рисунок 4.27 — Заполнение таблиц условных вероятностей вручную

После заполнении таблиц условных вероятностей производится проверка их заполненности, а также того, что сумма вероятностей для каждого значения должна быть равной 1. При ошибках заполнения таблиц условных вероятностей появятся предупреждающие сообщения.

После того как модель была обучена, то есть заполнены все таблицы условных вероятностей, можно приступать к оцениванию интенсивности пуассоновского процесса. Данные по которым нужно получить оценку также можно ввести тремя способами (рис. 4.28): можно синтезировать набор данных для того, чтобы проверить работу модели (при этом будут известны исходные интенсивности), можно ввести данные вручную, также можно выбрать файл формата csv или электронных таблиц Excel.

При выборе ввода данных вручную пользователь должен ввести данные об интервалах между эпизодами поведения в днях, по нажатию кнопки «Predict Behavior Rate» в соответствующем поле появляется оценка интенсивности поведения с заданной пользователем дискретизацией (рис. 4.29).

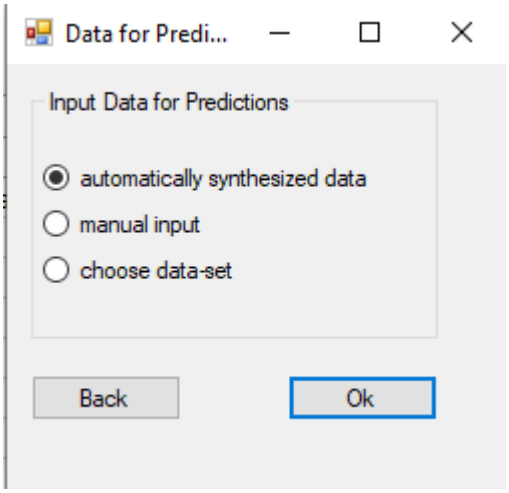

Рисунок 4.28 — Выбор данных для модели оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом

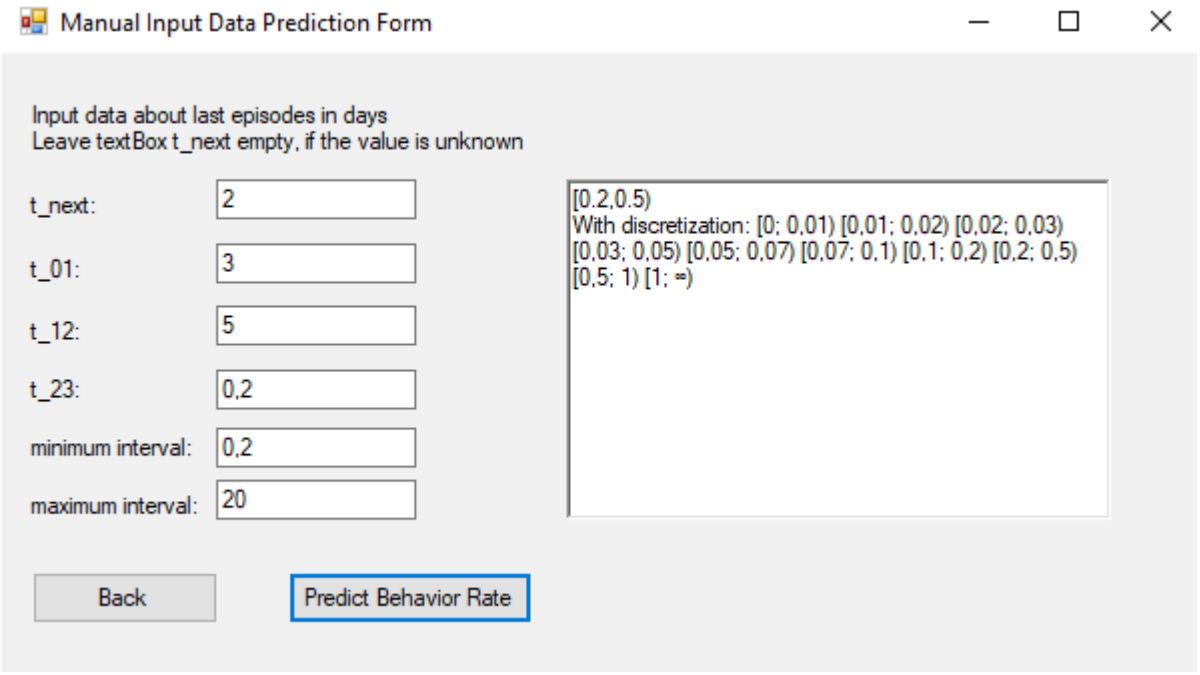

Рисунок 4.29 — Оценивание интенсивности пуассоновского процесса моделью при ручном вводе данных

При выборе синтетических данных пользователь снова должен указать все необходимые параметры (рис. 4.25), после этого результаты работы модели сохраняются в файл формата csv или таблиц Excel в указанном пользователем месте.

При выборе файла пользователь может выбрать файл с данными формата таблиц Excel или csv с помощью окна выбора файлов, после того как модель

получит оценки для всех записей, файл с результатами также сохраняется в формате csv или таблиц Excel в указанном пользователем месте.

4.4 ДАННЫЕ ДЛЯ АПРОБАЦИИ МОДЕЛЕЙ ОЦЕНИВАНИЯ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА

Сбор данных для апробации предложенных моделей также является довольно сложной задачей, так как необходимо найти источник данных о каком-либо процессе эпизодического поведения с известной интенсивностью.

В данной работе для апробации моделей были использованы синтетические данные, а также были собраны данные о постинге в социальных сетях.

4.4.1 Синтез данных последних эпизодах и рекордных интервалах пуассоновского процесса. Код для синтеза данных о последних эпизодах поведения написан на языке R. Программа генерирует «эпизоды поведения» согласно гамма-пуассоновской модели поведения. Для этого сначала случайным образом задаются 300 значений интенсивности, которые соответствуют значениям случайной величины с гамма-распределением, которое задается плотностью вероятности, имеющей вид: $f(x)=\begin{cases} x^{k-1}\dfrac{e^{-x/\theta}}{\theta^k\Gamma(k)}, x\geq 0 \\ 0, x<0 \end{cases}$, где $\Gamma(k)$ — гамма-функция Эйлера. При этом используются параметры $k=1.1$ (параметр формы) и $\theta=0.3$ (параметр масштаба), при таких параметрах большая часть значений меньше 0.5. Такие значения параметров отражают эмпирическую частоты некоторых реальных видов поведения [31].

Для заданных значений интенсивности синтезируется по 20 последовательностей точек, находящихся на расстоянии друг от друга, таким образом, что величины этих расстояний определены по экспоненциальному закону распределения со степенью равной соответствующему значению интенсивности. Далее, учитывая промежуток исследования (в данном случае 365 дней), для каждой последовательности берутся значения длин интервалов между последним

эпизодом и предпоследним, предпоследним и предпредпоследним (три правые точки на исследуемом промежутке). Также рассчитываются минимальное и максимальное значения интервалов между точками на выделенном промежутке. Кроме того, учитывается интервал между последней точкой (эпизодом), входящей в исследуемый промежуток, и ближайшей к ней точкой, не принадлежащей этому промежутку («следующий» эпизод), а также интервал между последней точкой (эпизодом) и точкой окончания сбора данных (конец исследуемого периода). Итоговый обучающий набор, полученный таким образом, содержит 6000 записей. Важно, что при этом задано исходное значение интенсивности. Это делает возможным его сравнение с оценкой модели.

Для синтеза ответов респондентов, то есть учитывая «зашумленность» ответов респондентов по отношению к настоящим данным о последних эпизодах и рекордных интервалах пуассоновского процесса используется следующий метод. На промежутках, зависящих от каждого из интервалов, выбираем, используя равномерное распределение, случайные величины так, чтобы выполнялись следующие условия:

- отклонение между окончанием сбора данных и последним эпизодом составляет не более четверти;
- между последними двумя эпизодами — не более половины;
- отклонение от интервала между предпоследним и предпредпоследним эпизодами максимум вдвое;
- отклонение от значений минимального и максимального интервала — не более четверти.

Такой подход может быть обоснован тем, что чем раньше эпизод, тем сложнее его вспомнить, а рекордные интервалы вспоминаются достаточно легко.

После того, как данные синтезированы значения непрерывных величин (временные интервалы и интенсивность) могут быть отнесены к определенному интервалу дискретизации непрерывных величин.

Важно, что полученные таким образом наборы данных содержат исходное значение интенсивности. Это делает возможным его сравнение с оценкой модели.

При рассмотрении работы моделей на синтетических данных берутся следующие дискретизации непрерывных величин: $\lambda^{(1)} = [0;0.01)$, $\lambda^{(2)} = [0.01;0.03)$, $\lambda^{(3)} = [0.03;0.05)$, $\lambda^{(4)} = [0.05;0.1)$, $\lambda^{(5)} = [0.1;0.2)$, $\lambda^{(6)} = [0.2;0.5)$, $\lambda^{(7)} = [0.5;1)$, $\lambda^{(8)} = [1;\infty)$ — для интенсивности; $t^{(1)} = \left[0;0.1\right)$, $t^{(2)} = \left[0.1;0.5\right)$, $t^{(3)} = \left[0.5;1\right)$, $t^{(4)} = \left[1;7\right)$, $t^{(5)} = \left[7;14\right)$, $t^{(6)} = \left[14;30\right)$, $t^{(7)} = \left[30;180\right)$, $t^{(8)} = \left[180;365\right)$, $t^{(9)} = \left[365;\infty\right)$ — для временных интервалов (единица измерения — 1 день).

4.4.2 Сбор данных из социальной сети ВКонтакте. В социальной сети ВКонтакте [4] у каждого пользователя есть своя страница, по желанию пользователя доступ к странице может быть ограничен. На каждой странице есть так называемая «стена», на этой «стене» пользователь может публиковать свои записи (посты), делать репосты записей других пользователей. Другие пользователи, если владелец «стены» не установил дополнительных ограничений, также могут оставлять записи на его «стене». Также пользователи могут добавлять в друзья других пользователей, выкладывать фотографии, аудио и видеозаписи, посылать друг другу сообщения, совершать видеозвонки, изменять информацию о себе, создавать сообщества, вступать в сообщества, отмечать «лайками», понравившиеся записи или фотографии, и многое другое. Все эти действия можно рассматривать как различные типы поведения. Ограничимся таким видом поведения как постинг на своей «стене».

Сбор данных осуществлялся посредством специально созданной программы на языке C#, для работы с ВКонтакте был использован API ВКонтакте [5].

Для доступа к стене пользователя применялся метод wall.get, предоставляемый API, с помощью этого метода можно получить информацию о 100 последних записях на «стене» пользователя, при этом можно отфильтровать записи таким образом, чтобы они принадлежали именно владельцу «стены». На

использование этого метода есть ограничение на 5000 вызовов в сутки, поэтому при работе с программой пользователю следует это учитывать.

В зависимости от настроек приватности пользователь может видеть полную информацию о владельце, имеющуюся на странице, и ограниченную общедоступными данными. Тот же принцип действует и в отношении API [5]. Поэтому для работы с программой требуется авторизация пользователя перед началом работы.

У каждого пользователя есть уникальный идентификатор по номеру (номера присваиваются в порядке создания новых страниц пользователями), также у пользователей есть возможность создать свой уникальный идентификатор и использовать его в написании адреса своей страницы.

Программа в качестве исходных данных о пользователе принимает или идентификатор по номеру страницы или идентификатор, созданный пользователем.

Данные о постинге можно получить как отдельно для каждого пользователя, так и сразу по списку пользователей, сохраненном в форматах .xls, .xlsx, .csv, также можно указать диапазон номеров страниц, по которым будут собираться данные. Кроме того, нужно указать период, за который будут собраны данные.

Результаты для отдельного пользователя ВКонтакте выводятся непосредственно в интерфейсе программы, также есть возможность сохранить их в форматах .xls, .xlsx, .csv, для списка или диапазона пользователей данные сразу сохраняются в файле, в выбранном пользователе формате.

Программа собирает из социальной сети ВКонтакте о каждом, указанном пользователе, если его или ее страница является открытой, следующие сведения:

- время публикации последнего поста на своей «стене» за определенный период времени;
- время публикации предпоследнего поста на своей «стене» за определенный период времени;

- время публикации предпредпоследнего поста на своей «стене» за определенный период времени;
- количество опубликованных за определенный период времени постов;
- время публикации первого поста на своей «стене» по истечении определенного периода времени.

Также программа позволяет рассчитать интенсивность постинга как отношение количества эпизодов постинга за период, определенный пользователем, к количеству временных единиц, также определенных пользователем (пользователь может выбрать секунды, минуты, часы или дни).

Время публикации записей сохранено в формате unixtime. Этот формат определяется как количество секунд, прошедших с полуночи (00:00:00 UTC) 1 января 1970 года [228].

Ниже представлен пример работы программы. После авторизации выберем в качестве периода июнь 2020 года, день в качестве единица измерения (см. рис. 4.30). В качестве примера возьмем диапазон страниц с идентификаторами о 1 до 100. Результаты сохраним в формате .xlsx.

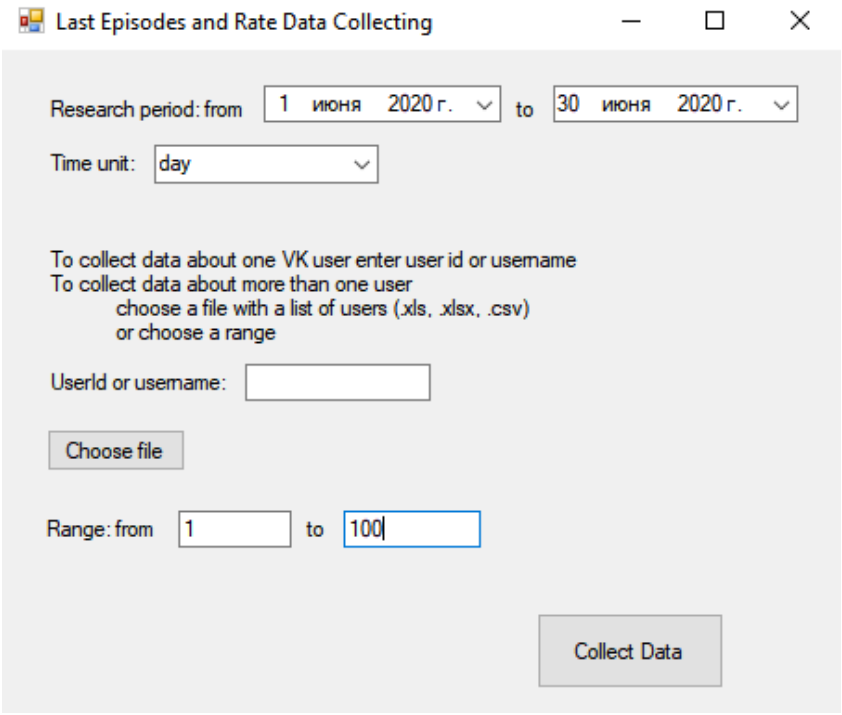

Рисунок 4.30 — Интерфейс программы для сбора сведений о постинге из социальной сети ВКонтакте [45]

На рис. 4.31 представлено начало полученного файла с результатами, строки с нулями означают, что за указанный период не было публикации постов, также пропущены некоторые идентификаторы, это значит, что доступ к страницам с этими идентификаторами закрыт.

UserInfoJun2020-100.xlsx *

	A	B	C	D	E	F	G	H	I
1	id	next	last	pre-last	pre-pre-last	min	max	rate	n
2	5	0	0	0	0	0	0	0	0
3	6	0	0	0	0	0	0	0	0
4	7	1594055140	0	0	0	0	0	0	0
5	8	0	0	0	0	0	0	0	0
6	10	0	0	0	0	0	0	0	0
7	11	0	0	0	0	0	0	0	0
8	14	1593788916	1593555664	1593433730	1593343364	90366	174683	0,129032258	4
9	15	0	1592931977	0	0	0	0	0,032258065	1
10	17	1593561907	1593439318	1593396226	1593348915	21661	603417	0,677419355	21
11	18	0	0	0	0	0	0	0	0

Рисунок 4.31 — Файл с результатами сбора данных о постинге в социальной сети ВКонтакте [45]

Был собран набор данных за декабрь 2019-го года о 6556 пользователях, 4590 записей было использовано для обучения моделей (70% от выборки), и 1966 — для тестирования (30% от выборки).

При рассмотрении работы моделей на данных, собранных из социальной сети ВКонтакте, берутся следующие дискретизации непрерывных величин: $\lambda^{(1)}=[0;0.1)$, $\lambda^{(2)}=[0.1;0.2)$, $\lambda^{(3)}=[0.2;0.3)$, $\lambda^{(4)}=[0.3;0.5)$, $\lambda^{(5)}=[0.5;1)$, $\lambda^{(6)}=[1;\infty)$ — для интенсивности; $t^{(1)}=[0;0.05)$, $t^{(2)}=[0.05;0.1)$, $t^{(3)}=[0.1;0.5)$, $t^{(4)}=[0.5;1)$, $t^{(5)}=[1;7)$, $t^{(6)}=[7;10)$, $t^{(7)}=[10;20)$, $t^{(8)}=[20;\infty)$ — для временных интервалов.

4.4.3 Сбор данных из социальной сети Instagram*. Сайт социальной сети Instagram* входил в число десяти наиболее популярных в России и является одним из наиболее посещаемых в мире [217]. В качестве исследуемого процесса будем рассматривать публикацию пользователями постов. Отметим, что речь идет именно о постах, публикации в «сторис» (временные публикации) не учитываются.

* Организация, запрещенная на территории РФ и признанная экстремистской.

На данный момент публикация постов является основным видом активности в этой социальной сети. Пост представляет собой фотографию, видео или несколько фотографий в виде галереи, время публикации поста фиксируется в формате unixtime [228] (то есть время представляется как прошедшее количество секунд с 1-го января 1970-го года до публикации). Таким образом, можно отследить какое число постов было опубликовано за интересующий нас период, чтобы вычислить интенсивность процесса.

На языке C# была написана программа для сбора данных из Instagram*. С помощью языка запросов GraphQL [142] она получает сведения о постах пользователя в формате JSON. В исследовании рассматривался 2019-ый год.

По id пользователя программа определяет время публикации трех его последних постов за 2019-й год, время публикации первого поста начиная с 1 января 2020-го, наибольший и наименьший интервалы между публикациями за 2019-й год и общее число постов за год. Также программа вычисляет значение интенсивности публикации постов, как отношение числа постов за исследуемый период к числу дней в этом периоде.

Был собран набор данных о 5608 пользователях, 3926 записей было использовано для обучения моделей (70% от выборки), и 1682 — для тестирования (30% от выборки).

При рассмотрении работы моделей на данных, собранных из социальной сети Instagram*, берутся следующие дискретизации непрерывных величин: $\lambda^{(1)} = [0; 0.002)$, $\lambda^{(2)} = [0.002; 0.01)$, $\lambda^{(3)} = [0.01; 0.03)$, $\lambda^{(4)} = [0.03; 0.05)$, $\lambda^{(5)} = [0.05; 0.1)$, $\lambda^{(6)} = [0.1; 0.2)$, $\lambda^{(7)} = [0.2; 0.5)$, $\lambda^{(8)} = [0.5; 1)$, $\lambda^{(9)} = [1; \infty)$ — для интенсивности; $t^{(1)} = \left[0; 0.1\right)$, $t^{(2)} = \left[0.1; 0.5\right)$, $t^{(3)} = \left[0.5; 1\right)$, $t^{(4)} = \left[1; 7\right)$, $t^{(5)} = \left[7; 14\right)$, $t^{(6)} = \left[14; 30\right)$, $t^{(7)} = \left[30; 180\right)$, $t^{(8)} = \left[180; 365\right)$, $t^{(9)} = \left[365; \infty\right)$ — для временных интервалов.

* Организация, запрещенная на территории РФ и признанная экстремистской.

4.4.4 Опросный инструментарий о последних постах в Instagram*. Для сбора данных респондентов был разработан опросный инструментарий на основе инструментов Google Сайты, Google Формы и Google Таблицы. Респондент заходит на сайт [20] отвечает на вопросы, после этого введенные сведения через дополнительную форму помещаются в таблицу.

Опросный инструментарий содержит следующие вопросы: имя пользователя, сведения об интервалах между последними тремя эпизодами публикации постов и данные о наибольшем и наименьшем интервалах между публикациями за год.

Ответы об интервалах между последними эпизодами публикации постов могут быть даны тремя способами:

- ввести дату и время (время необязательно, см. рис. 4.32);
- ввести интервал в выбранных единицах времени (рис. 4.33);
- ввести ответ в текстовое поле в свободной форме.

Ответы о минимальном и максимальном интервалах могут быть даны в свободной форме или с помощью ввода интервала в выбранных единицах времени (рис. 4.32).

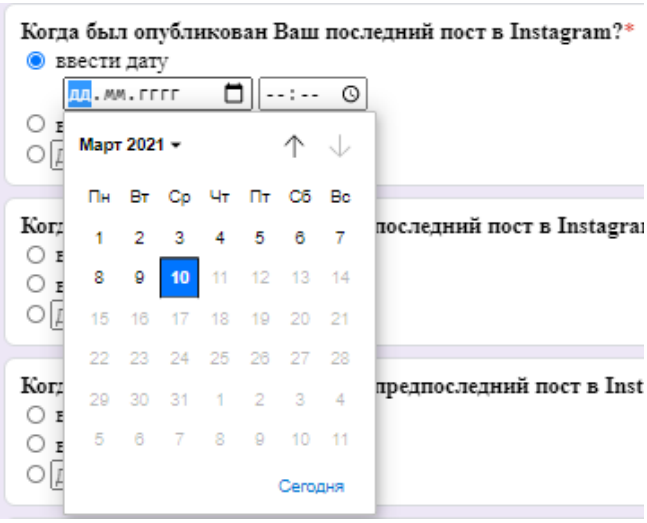

Рисунок 4.32 — Ввод даты публикации последнего поста [57]

* Организация, запрещенная на территории РФ и признанная экстремистской.

С помощью данного опросного инструментария были собраны данные о 92 пользователях Instagram*.

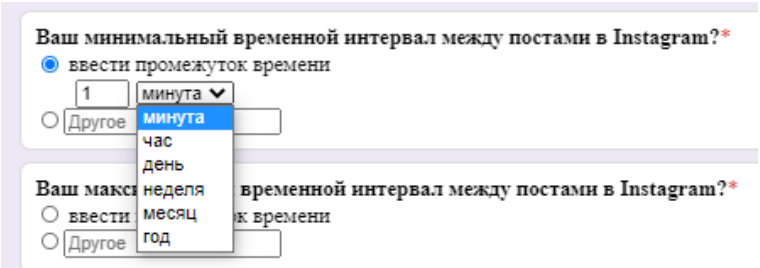

Рисунок 4.33 — Ввод временного интервала между публикацией постов [57]

В таблице 4.4 приведены примеры полученных ответов от респондентов.

Таблица 4.4 — Примеры ответов респондентов

№	t_1	t_2	t_3	t_{min}	t_{max}
1	1+month	3+month	не помню	2+week	6+month
2	4+hour	3+day	5+day	1+day	2+week
3	2/15/2021	2/14/2021	2/12/2021	1 день	2 недели
4	позавчера	месяц назад	полгода назад	3 минуты	4 месяца

Для того, чтобы привести эти данные к единому виду, а также для того, чтобы определить значение интенсивности для каждого пользователя на языке C# была написана специальная программа.

4.5 АПРОБАЦИЯ МОДЕЛЕЙ ОЦЕНИВАНИЯ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА

Собранные данные случайным образом разделялись на обучающую (80%) и тестовую (20%) выборки. Данные, собранные с помощью опросного инструментария о последних постах в Instagram*, полностью были использованы в качестве тестовой выборки для модели оценивания интенсивности пуассоновского процесса со скрытыми переменными.

4.5.1 Апробация модели оценивания интенсивности пуассоновского процесса со скрытыми переменными. Рассмотрим работы модели на

* Организация, запрещенная на территории РФ и признанная экстремистской.

синтетических данных. Таблица 4.5 представляет собой матрицу ошибок, полученную при работе модели на тестовой выборке.

Основные метрики качества получились следующие: BIC для тестовой выборки -47114.69, для учебной -90428.23; точность (acc.) 0.572, 95% доверительный интервал (0.5496; 0.5936); средняя точность 0.893; точность 0.499; полнота 0.474; F1-мера 0.486; каппа Коэна 0.466; при десятиклассовой кросс-валидации оценка ожидаемых потерь (expected loss) 10.

Таблица 4.5 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса со скрытыми переменными для синтетических данных

		Оценка интенсивности							
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$
Исходное значение интенсивности	$\lambda^{(1)}$	6	13	5	0	0	0	0	0
	$\lambda^{(2)}$	12	92	29	22	3	0	0	0
	$\lambda^{(3)}$	0	39	30	57	23	1	0	0
	$\lambda^{(4)}$	0	10	20	100	87	3	0	0
	$\lambda^{(5)}$	0	0	7	45	198	110	0	0
	$\lambda^{(6)}$	0	0	0	4	83	411	152	0
	$\lambda^{(7)}$	0	0	0	0	0	54	271	24
	$\lambda^{(8)}$	0	0	0	0	0	1	44	24

Далее рассмотрим работу модели на данных из социальных сетей, стоит учесть, что поскольку с помощью программного обеспечения были собраны точные данные об интенсивности публикации постов, «зашумленные» данные были синтезированы автоматически в соответствии с предположениями, использованными в разделе 4.4.1.

На данных из ВКонтакте были получены следующие результаты: BIC для тестовой выборки -38312.94, для учебной -69808.54; точность (acc.) 0.379, 95% доверительный интервал (0.3574; 0.4011); средняя точность 0.793; точность 0.355; полнота 0.342; F1-мера 0.348; каппа Коэна 0.217; при десятиклассовой кросс-

валидации оценка ожидаемых потерь 11.35. Матрица ошибок представлена таблицей 4.6.

На данных из Instagram* получились следующие показатели: BIC для тестовой выборки -47101.36, для учебной -74184.36; точность (acc.) 0.47, 95% доверительный интервал (0.4457; 0.4943); средняя точность 0.882; точность 0.515; полнота 0.503; F1-мера 0.509; каппа Коэна 0.382; при десятиклассовой кросс-валидации оценка ожидаемых потерь 10.66. Матрица ошибок представлена таблицей 4.7.

Таблица 4.6 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса со скрытыми переменными для данных из ВКонтакте

		Оценка интенсивности					
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$
Исходное значение интенсивности	$\lambda^{(1)}$	142	179	17	23	17	2
	$\lambda^{(2)}$	99	326	59	76	37	8
	$\lambda^{(3)}$	14	114	56	82	53	9
	$\lambda^{(4)}$	6	60	49	82	71	12
	$\lambda^{(5)}$	3	18	19	70	103	25
	$\lambda^{(6)}$	1	3	3	12	63	26

Таблица 4.7 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса со скрытыми переменными для данных из Instagram*

		Оценка интенсивности								
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$
Исходное значение интенсивности	$\lambda^{(1)}$	80	0	0	0	0	0	0	0	0
	$\lambda^{(2)}$	22	40	8	4	0	2	1	0	0
	$\lambda^{(3)}$	0	8	52	29	26	10	5	2	1
	$\lambda^{(4)}$	0	0	24	36	45	21	9	3	2
	$\lambda^{(5)}$	0	0	18	37	79	68	23	7	1
	$\lambda^{(6)}$	0	0	5	20	70	89	82	6	1
	$\lambda^{(7)}$	0	0	4	5	28	67	215	55	11
	$\lambda^{(8)}$	0	0	0	2	1	6	55	115	40
	$\lambda^{(9)}$	0	0	0	0	0	1	10	37	76

* Организация, запрещенная на территории РФ и признанная экстремистской.

С помощью опросного инструментария о последних постах в Instagram* (см. раздел 4.4.4) были собраны данные о 92 пользователях. Так как выборка имеет небольшой объем, то для обучения модели были использованы синтетические данные: по гамма-распределению были сгенерированы 500 значений интенсивностей, для каждого из этих значений было создано по 20 «респондентов», итоговый обучающий набор данных содержит 10000 записей.

Область значений гамма–распределеной величины λ была разбита на интервалы $\lambda^{(1)} = [0; 0.002)$, $\lambda^{(2)} = [0.002; 0.01)$, $\lambda^{(3)} = [0.01; 0.03)$, $\lambda^{(4)} = [0.03; 0.06)$, $\lambda^{(5)} = [0.06; 0.1)$, $\lambda^{(6)} = [0.1; 0.2)$, $\lambda^{(7)} = [0.2; 0.5)$. Отметим, что первый интервал ($\lambda^{(1)}$) подобран таким образом, чтобы в него входили те пользователи Instagram*, у которых нет публикаций за год.

Таблица 4.8 представляет собой матрицу ошибок, в которой строки обозначают истинные значения фактора, а столбцы — значения интенсивности, определенные моделью.

Результаты работы модели по основным метрикам качества получились следующими: BIC на обучающей выборке -141583.1, на тестовой -15551.2; точность (acc.) 0.355, средняя точность 0.839, точность 0.373, полнота 0.378, F1-мера 0.364, каппа Коэна 0.227 (для сравнения показатели на тех же данных исходной модели: BIC на обучающей выборке -106287.1, на тестовой -14259.8; точность (acc.) 0.254, средняя точность 0.813, точность 0.313, полнота 0.300, F1-мера 0.349, каппа Коэна 0.117). По полученным результатам было показано статистически достоверное улучшение точности на 6% и средней точности на 2.6% по сравнению с предложенным ранее алгоритмом: 95% доверительные интервалы для точности и средней точности, построенные методом бутстрап для 5000 репликаций, составляют (0.1972; 0.2535) и (0.7993; 0.8134) для исходной модели, а (0.2948; 0.4510) и (0.8237; 0.8627) — для предложенной модели оценивания интенсивности пуассоновского процесса со скрытыми переменными.

* Организация, запрещенная на территории РФ и признанная экстремистской.

Довольно высокий показатель средней точности говорит о том, что данная модель может быть применена для оценивания интенсивности в тех областях, где невозможно оценить этот параметр прямыми методами, а данные о процессе поведения, могут быть получены только с помощью опроса респондентов.

Таблица 4.8 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса со скрытыми переменными для данных пользователей Instagram*

		Оценка интенсивности							
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$
Исходное значение интенсивности	$\lambda^{(1)}$	8	0	0	0	0	0	0	0
	$\lambda^{(2)}$	2	3	2	1	0	0	0	0
	$\lambda^{(3)}$	0	6	6	3	2	1	0	0
	$\lambda^{(4)}$	0	2	4	4	2	0	0	0
	$\lambda^{(5)}$	0	0	5	6	3	2	0	0
	$\lambda^{(6)}$	0	0	0	3	3	2	1	0
	$\lambda^{(7)}$	0	0	1	3	0	0	2	0
	$\lambda^{(8)}$	0	0	0	0	0	0	0	0

4.5.2. Апробация модели оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом. Рассмотрим работы модели оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом на синтетических данных. Таблица 4.9 представляет собой матрицу ошибок, полученную при работе модели на тестовой выборке.

BIC для тестовой выборки -44295.16, для учебной -79880.18; точность (acc.) 0.619, 95% доверительный интервал (0.5971;0.6402); средняя точность 0.905; точность 0.588; полнота 0.568; F1-мера 0.578; каппа Коэна 0.53; при десятиклассовой кросс-валидации оценка ожидаемых потерь 8.09 (для сравнения результаты исходной модели на тех же данных: BIC для тестовой выборки

* Организация, запрещенная на территории РФ и признанная экстремистской.

- 40178.08, для учебной -71400.84; точность (acc.) 0.611, 95% доверительный интервал (0.5892; 0.6324); средняя точность 0.903; точность 0.58; полнота 0.559; F1-мера 0.569; каппа Коэна 0.52).

Таблица 4.9 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом для синтетических данных

		Оценка интенсивности							
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$
Исходное значение интенсивности	$\lambda^{(1)}$	19	18	2	0	0	0	0	0
	$\lambda^{(2)}$	10	96	37	15	2	0	0	0
	$\lambda^{(3)}$	0	34	36	69	11	0	0	0
	$\lambda^{(4)}$	0	3	17	120	79	1	0	0
	$\lambda^{(5)}$	0	0	0	32	235	93	0	0
	$\lambda^{(6)}$	0	0	0	1	91	398	160	0
	$\lambda^{(7)}$	0	0	0	0	0	28	293	29
	$\lambda^{(8)}$	0	0	0	0	0	0	30	40

На данных из ВКонтакте были получены следующие результаты: BIC для тестовой выборки -35335.54, для учебной -62711.71; точность (acc.) 0.448, 95% доверительный интервал (0.4254; 0.4699); средняя точность 0.816; точность 0.444; полнота 0.4; F1-мера 0.421; каппа Коэна 0.303; при десятиклассовой кросс-валидации оценка ожидаемых потерь 9.76 (для сравнения результаты исходной модели на тех же данных: BIC для тестовой выборки -31470.19, для учебной - 55191.93; точность (acc.) 0.441, 95% доверительный интервал (0.4193; 0.4637); средняя точность 0.814; точность 0.429; полнота 0.399; F1-мера 0.413; каппа Коэна 0.295). Матрица ошибок представлена таблицей 4.10.

На данных из Instagram* получились следующие показатели: BIC для тестовой выборки -45352.94, для учебной -68905.34; точность (acc.) 0.528, 95% доверительный интервал (0.5041; 0.5524); средняя точность 0.895; точность 0.592;

* Организация, запрещенная на территории РФ и признанная экстремистской.

полнота 0.562; F1-мера 0.577; каппа Коэна 0.448; при десятиклассовой кросс-валидации оценка ожидаемых потерь 9 (для сравнения результаты исходной модели на тех же данных: BIC для тестовой выборки -41088.13, для учебной - 61703.63; точность (acc.) 0.527, 95% доверительный интервал (0.5026; 0.5508); средняя точность 0.895; точность 0.591; полнота 0.559; F1-мера 0.575; каппа Коэна 0.447). Матрица ошибок представлена таблицей 4.11.

Таблица 4.10 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом для данных из ВКонтакте

		Оценка интенсивности					
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$
Исходное значение интенсивности	$\lambda^{(1)}$	189	162	12	9	23	0
	$\lambda^{(2)}$	92	372	33	63	47	3
	$\lambda^{(3)}$	6	124	55	98	47	2
	$\lambda^{(4)}$	3	41	46	93	93	5
	$\lambda^{(5)}$	0	11	2	65	149	11
	$\lambda^{(6)}$	0	1	0	8	78	21

Таким образом показатели качества оценки в предложенных моделях оценивания интенсивности пуассоновского процесса выше по сравнению с исходными. Кроме того, использование предложенных методов позволяет снизить влияние неопределенности при оценивании интенсивности эпизодического поведения индивидов что немаловажно в областях, где данные получать сложно и дорого.

Таблица 4.11 — Матрица ошибок модели оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом для данных из Instagram*

		Оценка интенсивности								
		$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$	$\lambda^{(9)}$
Исходное значение интенсивности	$\lambda^{(1)}$	81	0	0	0	0	0	0	0	0
	$\lambda^{(2)}$	4	59	13	2	2	2	0	0	0
	$\lambda^{(3)}$	0	2	71	29	26	4	3	3	1
	$\lambda^{(4)}$	0	0	19	38	46	28	4	3	2
	$\lambda^{(5)}$	0	1	14	34	85	73	24	3	2
	$\lambda^{(6)}$	0	0	1	7	60	123	79	2	1
	$\lambda^{(7)}$	0	0	0	3	21	66	239	49	8
	$\lambda^{(8)}$	0	0	0	0	3	5	62	119	31
	$\lambda^{(9)}$	0	0	0	1	0	0	9	41	73

4.5.3. Апробация моделей оценивания интенсивности постинга в социальной сети, расширенной объективными данными о пользователе. Для обучения и тестирования моделей были собраны данные о пользователях социальной сети ВКонтакте: больше 9 тысяч для обучающей и тестовой выборки, были взяты только полные наблюдения без отсутствующих значений, то есть для каждого пользователя были известны количество друзей, пол, возраст, а также хотя бы 3 поста в течение года. После удаления пользователей с неполными данными в обучающей выборке осталось 1490 пользователей, а в тестовой 1472 пользователя. Стоит отметить, что из выборки также по этическим соображениям исключались пользователи младше 18 лет. Были выбраны следующие интервалы дискретизации: для интенсивности постинга — $[0;0.015)$, $[0.015;0.03)$, $[0.03;0.06)$, $[0.06;0.14)$, $[0.14;\infty)$, для временных интервалов — $[0;1)$, $[1;7)$, $[7;14)$, $[14;30)$, $[30;90)$, $[90;\infty)$ (единица измерения — день). Для возраста было взято 2 интервала: младше 35 лет

* Организация, запрещенная на территории РФ и признанная экстремистской.

и старше 35 лет, так как возраст до 35 лет в Российской Федерации считается возрастом молодежи.

Статистически объективные показатели действительно связаны с интенсивностью постинга (рис. 4.34).

Интенсивность постинга статистически значимо различается между мужчинами и женщинами (с поправкой на возраст): в рассматриваемой выборке женщины чаще публикуют посты. Для каждого интервала интенсивности постинга, кроме самого высокого, число друзей у мужчин и женщин статистически значимо отличается.

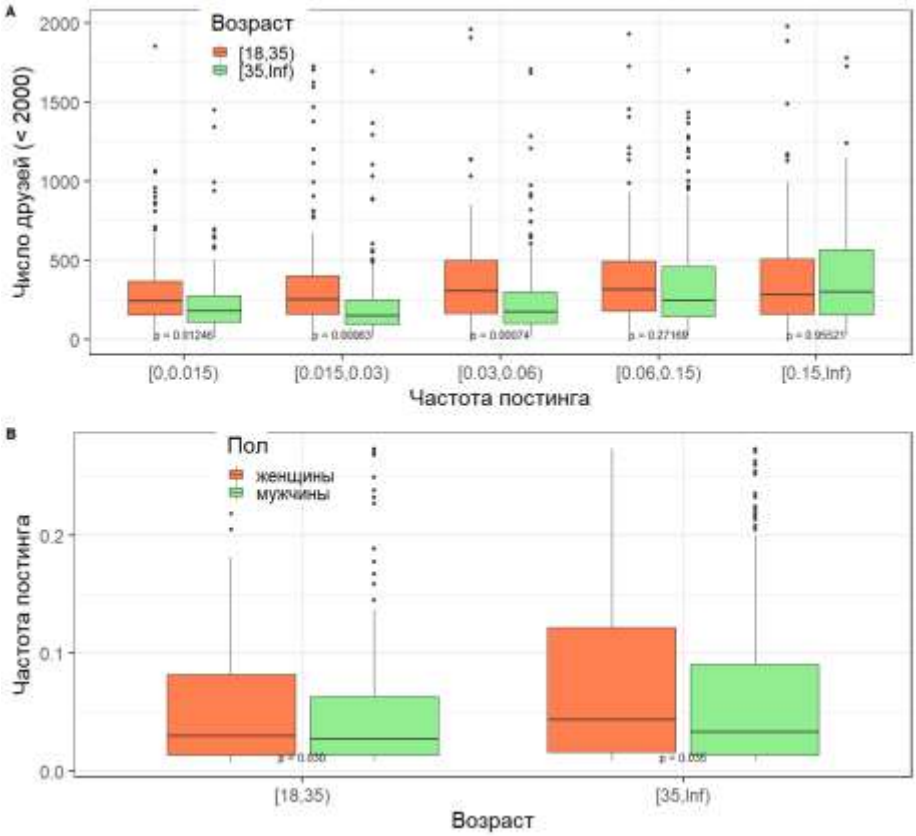

Рисунок 4.34 — Статистические характеристики тестовой выборки. Сверху: зависимость интенсивности постинга от числа друзей с группировкой по полу. Снизу: зависимость рассчитанной интенсивности постинга от пола с группировкой по возрасту [28]

Для моделей были вычислены значения информационных критериев, а также показатели качества классификации, также для сравнения они же были вычислены

для упрощенной модели (не содержащей узлов *age*, *sex*, *friends count*). В таблице 4.12 представлены значения информационных критериев (информационный критерий Акаике (AIC), Байесовский информационный критерий (BIC) и критерий отношения правдоподобия (loglik)) для трех моделей на двух выборках: обучающей и тестовой.

Далее были рассчитаны показатели качества моделей на тестовой выборке, показатели качества представлены в таблице 4.13. Все три модели оказались сопоставимы, что объясняется тем, что связи с личными характеристиками пользователя являются довольно слабыми (см. таблица 3.15).

Таблица 4.12 — Информационные критерии

Модель	тренировочная выборка			тестовая выборка		
	AIC	BIC	loglik	AIC	BIC	loglik
упрощенная	-9505	-9715	-9426	-9298	-9507	-9219
расширенная	-13542	-13924	-13398	-13223	-13604	-13398
с обученной структурой	-13510	-13818	-13394	-13196	-13503	-13079

Таблица 4.13 — Критерии качества

Модель	Точность	95% Доверительный интервал	Каппа	Средняя точность
упрощенная	0.475	(0.4491, 0.5007)	0.334	0.788
расширенная	0.481	(0.4552, 0.5069)	0.342	0.791
с обученной структурой	0.477	(0.4511, 0.5028)	0.336	0.789

4.6 ВНЕДРЕНИЕ МЕТОДОВ И АЛГОРИТМОВ ОБРАБОТКИ НЕОПРЕДЕЛЕННОСТИ ПРИ ОЦЕНИВАНИИ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА

Все положения, выносимые на защиту, были внедрены в научно-исследовательской работе СПб ФИЦ РАН № 0073-2019-0003 «Состояние и

перспективы развития информационного общества и цифровой экономики в России», в рамках которой решалась задача автоматизации сбора информации об эпизодическом поведении индивида и вычислении сводных характеристик такого поведения. В условиях ограниченности ресурсов особую роль играют предложенные соискателем модели учета неопределенности, сопутствующей самоотчетам респондентов.

Основная задача исследования состоит в повышении качества классификации при оценивании интенсивности пуассоновского процесса, моделирующего реализацию эпизодов поведения индивида, по данным, содержащим значительные неточности или ошибки, свойственные ситуации сбора самоотчетов респондентов об их поведении.

Для обработки неопределенности, возникающей при задании конца исследуемого периода, были предложены метод и алгоритм оценивания интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений, в основе которого лежит БСД с дополнительным интервалом между эпизодами процесса. Для подтверждения качества работы модели были использованы синтетические данные, в которых длина интервала между моментом интервью и последним эпизодом пуассоновского процесса зашумлена с использованием усеченного нормального распределения со следующими параметрами: математическое ожидание равно длине интервала между моментом интервью и последним эпизодом пуассоновского процесса, дисперсия равна удвоенной длине интервала между моментом интервью и последним эпизодом пуассоновского процесса. Было показано статистически достоверное улучшение средней точности (средняя точность равна 0.904) на 0.6% по сравнению с методом, предложенным ранее (средняя точность равна 0.898): 95% доверительный интервал для средней точности, построенный методом бутстрап для 20000 репликаций, составляет (0.8951; 0.8992) — для исходной модели оценивания интенсивности пуассоновского процесса и (0.9016; 0.9050) — для модели оценивания интенсивности пуассоновского процесса, обрабатывающей неопределенность задания конца исследуемого периода.

Для обработки некорректности полученной от респондентов информации, был предложен алгоритм оценивания интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, в основе которого лежит расширенная при помощи скрытых переменных байесовская сеть доверия. Были проведены численные эксперименты, и для разных моделей ошибки было получено статистически достоверное улучшение качества классификации на 1%. В таблице 4.14 представлено значение средней точности и 95% доверительный интервал, построенный методом бутстрап для 2000 репликаций.

Улучшение качества классификации на 1% важно при проведении популяционных исследований, так как в этом случае даже такое улучшение имеет значительный экономический эффект.

Таблица 4.14 — Сравнение средней точности оценки интенсивности пуассоновского процесса в исходной модели оценивания интенсивности пуассоновского процесса и модели со скрытыми переменными

	Исходный алгоритм оценивания интенсивности пуассоновского процесса		Алгоритм оценивания интенсивности пуассоновского процесса на основе модели со скрытыми переменными		Улучшение показателя
	95% доверительный интервал для средней точности	Средняя точность	95% доверительный интервал для средней точности	Средняя точность	
Равномерное распределение на отрезке от 0 до удвоенной длины истинного интервала	(0.8693; 0.8730)	0.872	(0.8755; 0.8823)	0.879	0.1%
Нормальное распределение с математическим ожиданием равным длине интервала и дисперсией распределения равной длине интервала	(0.8328; 0.8363)	0.834	(0.8692; 0.8757)	0.875	4.1%
Треугольное распределение на	(0.8726; 0.8763)	0.874	(0.8802; 0.8875)	0.887	1.3%

отрезке от половины длины до утроенной длины истинного интервала с модой распределения равной длине интервала					

Предложенные алгоритмы были реализованы в программном обеспечении. Программный модуль для работы с моделью оценивания интенсивности пуассоновского процесса со скрытыми переменными был внедрен в процесс консультирования клиентов ООО «Хоум Фитнес» для определения оптимального уровня физической нагрузки, соответствующего рекомендациям Всемирной Организации Здравоохранения (ВОЗ).

В качестве исходной задачи выступил запрос от организации на создание инструмента для оценки частоты физических упражнений в рамках рутинного анкетирования клиента и сравнение индивидуального показателя с предписаниями ВОЗ с целью обоснования рекомендаций по повышению физической активности клиента и мотивации к более частому использованию ресурсов организации.

Использование алгоритма обработки некорректности информации об эпизодах поведения, полученной от респондентов, при оценивании интенсивности поведения клиента оправдано тем, что кроме физической активности, которую можно отследить по абонементу, существует и повседневная деятельность, подразумевающая определенную физическую нагрузку.

Процесс построения режима физических нагрузок для клиента выглядит следующим образом. Клиента опрашивают о последних трех эпизодах физической активности средней или высокой интенсивности (определяется согласно ВОЗ [248]) длительностью более получаса, а также о минимальном и максимальном интервалах между эпизодами физической активности средней или высокой интенсивности. Вычисляется оценка частоты физической активности клиента за месяц и сопоставляется с рекомендациями ВОЗ [248]. Например, для взрослых людей от 18 до 64 лет рекомендуется не менее 150 – 300 минут в неделю заниматься физически активной деятельностью средней интенсивности с аэробной нагрузкой,

то есть значение рекомендуемой частоты физической активности за месяц должно находиться в пределах интервала [0.446; 0.893]. Если оценка частоты физической активности клиента оказывалась ниже, то рекомендовалось повысить частоту физических нагрузок. Специалистами ООО «Хоум Фитнес» строится подходящий для клиента режим физических нагрузок на основе доступных в фитнес-центре занятий.

Таким образом, предложенный инструментарий используется специалистами ООО «Хоум Фитнес» для поддержки принятия решений о рекомендуемом уровне физической нагрузки, а также в качестве доказательной базы при мотивации клиента. Результатом является меньшая нагрузка при анкетировании, так как информация о нескольких последних эпизодах поведения обычно легко припоминается, и построение более аргументированного плана физических нагрузок. Кроме того, использование опросного инструментария не требует специальной подготовки со стороны интервьюера.

Также модели со скрытыми переменными и методика сбора сведений о последних эпизодах в социальных сетях используются в учебном процессе факультета государственного и муниципального управления Северо-Западный институт управления, филиала Федерального государственного бюджетного образовательного учреждения высшего образования «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации» по программе второго высшего образования при проведении практических и теоретических занятий по дисциплине «Стратегия управления человеческими ресурсами». Методика сбора сведений о последних эпизодах постинга в социальных сетях используется при разработке опросных инструментариев для оценки рискообразующего поведения сотрудников, а оценки интенсивности такого поведения, рассчитанные с помощью моделей оценивания интенсивности пуассоновского процесса со скрытыми переменными, учитываются при принятии управленческих решений.

ВЫВОДЫ ПО ГЛАВЕ 4

В данной главе представлена реализация методов и алгоритмов, предложенных в третьей главе. Описаны архитектура и прототип комплекса программ для работы с моделями оценивания интенсивности пуассоновского процесса. В первом разделе описан модуль для работы с инструментом оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса, его использование позволяет обеспечить необходимую в практической задаче степень согласованности данных респондентов. Во втором разделе описан модуль для работы с моделью оценивания интенсивности пуассоновского процесса со скрытыми переменными, где скрытые данные соответствуют истинным значениям о последних эпизодах и рекордных интервалах пуассоновского процесса, использование данного модуля позволяет обрабатывать неопределенность, связанную с намеренными и ненамеренными искажениями ответов респондентов на вопросы о последних эпизодах их поведения. В третьем разделе описан модуль для работы с моделью оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом, при наличии ретроспективных данных его использование позволяет повысить качество классификации при оценивании интенсивности пуассоновского процесса. В четвертом разделе описаны процедуры синтеза и сбора данных для апробации моделей, составляющих основу предложенных методов и алгоритмов. Синтез данных для обучения моделей оценивания интенсивности пуассоновского процесса используется при невозможности обучить модели на реальных данных, как было сделано в разделе 4.5.1. При сборе данных из социальных сетей в качестве пуассоновского процесса рассматривалась публикация постов. Благодаря тому, что оценку интенсивности публикации постов в социальной сети довольно просто получить в явном виде, на данных из социальных сетей возможно оценить качество классификации предложенных моделей. В пятом разделе проведена апробация предложенных моделей на собранных данных. Было показано, что показатели качества классификации предложенных моделей статистически выше, чем у

исходной модели, описанной в 2.2. Также представлены результаты внедрении полученных в ходе исследования результатов в научно-исследовательской работе СПб ФИЦ РАН, ООО «Хоум Фитнес» и в СЗИУ РАНХиГС.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена научная задача обработки некоторых типов неопределенности данных при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений. Было повышено качество классификации при оценивании интенсивности пуассоновского процесса как математической модели эпизодического поведения индивида за счет разработки методов и алгоритмов обработки неопределенности данных, предоставляемых респондентами. Решенная задача имеет важное значение для совершенствования методов и алгоритмов, используемых в системном анализе, оптимизации, управлении, принятии решений и обработке информации, связанных с моделированием человеческого поведения и улучшением качества оценок его интенсивности.

Итоги исследования включают нижеперечисленные научные результаты:

— Разработаны метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса, в основе которых лежит расширенная байесовская сеть доверия с дополнительными узлами принятия решений. С помощью этого инструмента можно оценить, насколько согласована информация, полученная от респондентов, и дальше действовать в зависимости от целей исследования: например, исключить из рассмотрения данные, не удовлетворяющие определенному порогу согласованности.

— Разработан алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, в основе которого лежит использование скрытых переменных в байесовской сети доверия, отвечающих истинным длинам интервалов. Было показано, что использование в модели таких скрытых переменных позволяет улучшить показатели качества классификации по сравнению с предложенным ранее подходом.

— Разработаны метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского

процесса по ограниченному объему доступных наблюдений на основе байесовской сети доверия, при этом в модель на этапе обучения вводится вершина, характеризующая интервал между последним эпизодом пуассоновского процесса и эпизодом, произошедшим после окончания периода исследования. При наличии ретроспективных данных для исследования использование этой модели позволяет улучшить показатели качества классификации по сравнению с предложенным ранее подходом.

— Разработаны архитектура и прототип комплекса программ для работы с предложенными новыми методами и алгоритмами для их апробации, вычислительных экспериментов и решения практических задач.

По итогам исследования даны **рекомендации** по применению полученных результатов в научных исследованиях и прикладных задачах, в которых важно оценивать числовые характеристики поведения человека. Результаты, представленные в диссертации, применимы в качестве инструмента автоматизации оценивания одного из ключевых параметров эпизодического поведения — его интенсивности, с учетом неточности и некорректности данных самоотчетов респондентов. Результаты данного исследования предназначены для использования специалистами, изучающими поведение человека, в условиях ограниченности ресурсов. С помощью разработанных инструментов можно получить довольно высокое качество оценивания интенсивности поведения на основе небольшого количества данных.

Метод оценивания согласованности информации о поведении респондента можно использовать для того, чтобы отфильтровать сведения респондентов ненадлежащего качества. Алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, по ограниченному объему доступных наблюдений, на основе модели БСД со скрытыми переменными, можно использовать в тех случаях, когда особенно важно учитывать то, что респонденты могут предоставить неверные данные. Метод обработки неопределенности задания конца исследуемого периода при оценивании

интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений рекомендован для ретроспективных исследований.

Перспективами дальнейшей разработки тематики могут стать исследования, направленные на включение в модели оценивания интенсивности пуассоновского процесса дополнительных параметров, интересующих исследователя, а также учета дополнительных сведений, полученных от экспертов. Еще одним направлением может стать использование в методах оценивания интенсивности поведения синтезированных по данным моделей, а также более детальный подход к созданию синтетических данных для таких задач.

Полученные результаты соответствуют специальности 2.3.1 – «Системный анализ, управление и обработка информации, статистика».

СЛОВАРЬ ТЕРМИНОВ

Байесовская сеть доверия (БСД) — это вероятностная графическая модель, которая представляет собой ациклический направленный граф, его вершинами являются случайные элементы, входящие в модель, а ребра обозначают причинно-следственные связи между элементами. С ребрами сети связаны тензоры условной вероятности. [67, 68].

Дискретизация — разбиение области допустимых значений непрерывной величины на конечное число дизъюнктных интервалов [110, 144].

Интенсивность — отношение количества эпизодов процесса за период исследования к количеству временных единиц, составляющих этот период исследования [134].

Конец периода исследования (окончание периода исследования, момент сбора данных, момент интервью) — точка на временной оси, которая соответствует завершению периода исследования, в прикладных задачах соответствует моменту сбора данных от респондента [30, 72].

Матрица ошибок — таблица, в которой строках указаны действительные значения, а в столбцах — оценки, полученные моделью. Такая таблица дает общее представление о качестве классификации модели [17].

Некорректность данных — расхождение данных с истинными значениями. В диссертационном исследовании под некорректностью данных понимается некорректность данных, полученных вследствие ответов респондентов о последних эпизодах и рекордных интервалах их поведения.

Несогласованность данных — внутренняя противоречивость данных. В диссертационном исследовании под несогласованностью данных понимается несогласованность данных, полученных вследствие ответов респондентов о последних эпизодах и рекордных интервалах их поведения, то есть рассматриваются ситуации, когда респондент дает ответы, противоречащие друг другу.

Неопределенность — это отсутствие полной информации об интересующем объекте [204].

Ограниченный набор наблюдений — сверхкороткие, неполные и неточные данные о нескольких последовательных эпизодах и рекордных интервалах процесса.

Период исследования — заранее определенный и зафиксированный промежуток на временной оси, в течение которого исследуется поведение индивидов [30, 72].

Поведение — активность субъекта, проявляющаяся при его взаимодействии с окружающей средой. В диссертационном исследовании рассматривается эпизодическое поведение, то есть выражающееся многократными, разделенными по времени активностями [134].

Последовательные эпизоды — эпизоды, следующие один за другим.

Пуассоновский процесс — ординарный поток однородных событий, для которого число событий в интервале T не зависит от чисел событий в любых интервалах, не пересекающихся с T, и подчиняется распределению Пуассона. В теории случайных процессов описывает количество наступивших случайных событий, происходящих с постоянной интенсивностью [12].

Рекордные интервалы — минимальный и максимальный интервалы между последовательными эпизодами процесса за исследуемый период [71, 72].

Респондент — субъект, принимающий участие в опросе, анкетировании [71, 72].

Самоотчет респондента — ответы в анкете на определенные вопросы или результаты проведения интервью [71, 72].

Скрытые переменные — переменные, которые не могут быть измерены в явном виде, а могут быть только выведены через математические модели с использованием наблюдаемых переменных [67, 185].

Эпизод поведения — это активность, произошедшая в определенное время, у которой можно определить начало и конец [134].

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

БСД	Байесовские сети доверия
AIC	Информационный критерий Акаике (Akaike Information Criterion)
BIC	Информационный критерий Байеса (Bayesian Information Criterion)
EF	Алгоритм разделения области на интервалы равные по частоте (Equal Frequency)
EM	Максимизация ожидания (Expectation Maximization)
EW	Алгоритм разделения области на интервалы равные по длине (Equal Width)
FN	Ложно-отрицательные наблюдения (False Negative)
FP	Ложно-положительные наблюдения (False Positive)
MLE	Метод максимального правдоподобия (Maximum Likelihood Estimate)
TN	Истинно-отрицательные наблюдения (True Negative)
TP	Истинно-положительные наблюдения (True Positive)

СПИСОК ЛИТЕРАТУРЫ

1. Бабиков, В.М. Некоторые аспекты применения байесовых сетей для оценки надежности автоматизируемых человеко-машинных систем / В.М. Бабиков // Труды международной научно-практической конференции «Передовые информационные технологии, средства и системы автоматизации и их внедрение на российских предприятиях» (AITA-2011). — Москва, 2011. — С. 266–276.

2. Белозерский, А.Ю. Использование аппарата нечетких байесовых сетей для оценки инновационных рисков / А.Ю. Белозерский, Т.В. Какатунова, И.В. Иванова // Транспортное дело России. — 2011. — № 2. — С. 43–46.

3. Бычков, Е.Д. Нечеткая байесовская сеть в модели качества обслуживания услуг сети связи / Е.Д. Бычков, Б.К. Сагинова, Н.Н. Нарутта // Россия молодая: передовые технологии – в промышленность! — 2013. — № 1. — С. 202–205.

4. ВКонтакте [Электронный ресурс]. — Режим доступа: http://www.vk.com.

5. ВКонтакте. Описание методов API [Электронный ресурс]. — Режим доступа: https://vk.com/dev/methods.

6. Гаврилина, В.Ф. Подход байеса в управлении факторами техногенных рисков при кредитовании в коммерческом банке / В.Ф. Гаврилина // Проблемы анализа риска. — 2012. — Т. 9. — № 3. — С. 68–79.

7. Журкина, Л.С. Факторы, определяющие поведение покупателей в сети Интернет / Л.С. Журкина, Ю.А. Уханова, А.Ф. Никишин, Т.В. Панкина //Современные научные исследования и инновации. — 2015. — № 6–4 (50). — С. 24–26.

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

9. Козлов, А.А. Рискованные формы поведения и уровень социального функционирования ВИЧ-позитивных потребителей инъекционных наркотиков: прогностические модели / А.А. Козлов, Э.П. Станько, С.А. Игумнов // Медицинская психология в России. — 2018. — Т. 10. — 1 (48). — С. 1–22.

10. Кузнецов, А.Б. Методика диагностирования автоматизированных систем управления сложными объектами с использованием априорной информации / А.Б. Кузнецов, Н.А. Осипов, И.В. Дорожко // Известия высших учебных заведений. Приборостроение. — 2013. — Т. 56. — № 1. — С. 18–26.

11. Маклакова, Г.Г. Система оценки качества услуг телекоммуникационной сети дистанционного обучения на основе байесовских сетей доверия / Г.Г. Маклакова // Новые компьютерные технологии. — 2008. — Т. 6. — № 1 (6). — С. 72–73.

12. Математическая энциклопедия / Главный редактор И. М. Виноградов. — М.: «Советская энциклопедия», 1979. — Т. 4. — 1104 с.

13. Микони, С.В. Квалиметрия моделей и полимодельных комплексов / С.В. Микони, Б.В. Соколов, Р.М. Юсупов. —2018. — М.: РАН. —314 с.

14. Морозков, А.А. Обучение глобальной структуры байесовских сетей доверия на основе роевых биологических метаэвристик/ А.А. Морозков, А.А. Фильченков // Международная конференция по мягким вычислениям и измерениям. — 2015. — Т. 1. — № Секции 1–3. — С. 49–52.

15. Мусина, В.Ф. Байесовские сети доверия как вероятностная графическая модель для оценки медицинских рисков / В.Ф. Мусина // Труды СПИИРАН. — 2013. — Вып. 24. — С. 135–151.

16. Мусина, В.Ф. Байесовские сети доверия как вероятностная графическая модель для оценки экономических рисков / В.Ф. Мусина // Труды СПИИРАН. — 2013. — Вып. 25. — С. 235–254.

17. Мухамедиев, Р.И. Таксономия методов машинного обучения и оценка качества классификации и обучаемости / Р.И. Мухамедиев, Е.Л. Мухамедиева, Я.И. Кучин // Cloud of science. — 2015. — Т. 2. — №3. — С. 359–378.

18. Мхитарян, С.В. Применение статистических методов для анализа и моделирования поведения клиентов / С.В. Мхитарян // Инновации в гражданской авиации. — № 1. — 2016.

19. Некрасов, С.Н. Комбинированный метод оценки навигационной безопасности при плавании по внутренним водным путям / С.Н. Некрасов, А.А. Прохоренков // Вестник государственного университета морского и речного флота им. адмирала С.О. Макарова. — 2011. — № 1. — С. 106–108.

20. Опросный инструментарий. [Электронный ресурс]. — Режим доступа: https://sites.google.com/view/instagrampostsquestionnary/.

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

22. Плавинский, С.Л. Сексуальное поведение, венерические болезни и гетеросексуальная эпидемия ВИЧ-инфекции — некоторые результаты математического моделирования / С.Л. Плавинский, А.Н. Баринова, К.И. Разнатовский // Российский семейный врач. — 2007. — Т. 11. — № 3. — С. 30–38.

23. Потрясаев, С.А. Полимодельный комплекс мобильной сервисной системы, предназначенной для обслуживания воздушных судов / С.А. Потрясаев, А.Л. Ронжин, Б.В. Соколов, В.Ю.-Д. Джао, П.В. Степанов, М.М. Стыскин // Информатизация и связь. — 2020. — № 6. — С. 113–118.

24. Смирнов, А.В. Базовый сценарий интеллектуальной поддержки принятия решений на основе моделей жизни пользователей в цифровой среде / А. В. Смирнов, Т. В. Левашова, М. В. Петров // Информационно-управляющие системы. — 2021. — № 4(113). – С. 47– 60. — Doi: 10.31799/1684-8853-2021-4-47-60

25. Смирнов, В.А. Поиск неисправностей в бортовых системах управления в процессе приемочного контроля / В.А. Смирнов // Информационно-управляющие системы. — 2013. — № 2 (63). — С. 24–28.

26. Соколов, С.Н. Моделирование поведения пользователей интернет-ресурсов на основе смеси цепей Маркова / С.Н. Соколов // Естественные и технические науки. — 2009. — № 5. — С. 302–305.

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

28. Столярова, В.Ф. Модель для оценки частоты публикации постов в онлайн социальной сети по неполным данным с учетом объективных детерминант поведения / В.Ф. Столярова, А.В. Торопова, А.Л. Тулупьев // Нечеткие системы и мягкие вычисления. — 2021. — Т. 16. — № 2. — С. 77–95. — Doi: 10.26456/fssc81.

29. Студенников, К.О. Об одном подходе к управлению информационными рисками на основе коэффициентов чувствительности/ К.О. Студенников, В.Н. Лопин // Информация и безопасность. — 2013. — Т. 16. — № 2. — С. 219–222.

30. Суворова, А.В. Моделирование социально-значимого поведения по сверхмалой неполной совокупности наблюдений / А.В. Суворова // Информационно-измерительные и управляющие системы. — 2013. — №9. — т. 11. — С. 34–38.

31. Суворова, А.В. Синтез структур байесовской сети доверия для оценки характеристик рискованного поведения / А.В. Суворова, А. Л. Тулупьев // Информационно-управляющие системы. — 2018. — № 1. — С. 116–122. — DOI: 10.15217/issn1684-8853.2018.1.116.

32. Суворова, А.В. Байесовские сети доверия в задачах оценивания интенсивности рискованного поведения / А.В. Суворова, А. Л. Тулупьев, А.В. Сироткин // Нечеткие системы и мягкие вычисления. — 2014. — Т. 9. — № 2. — С. 115–129.

33. Судаков, К.А. Анализ и прогноз поведения потребителей / К.А. Судаков //Управление продажами. — 2015. — № 1. — С. 74–80.

34. Торопова, А.В. Байесовские сети доверия: инструменты и использование в учебном процессе / А.В. Торопова // Компьютерные инструменты в образовании. — №4. — 2016. — С. 43–53.

35. Торопова, А.В. Анализ согласованности данных в модели оценки интенсивности социально-значимого поведения / А.В. Торопова // Актуальные направления научных исследований XXI века: теория и практика. — Том 3. — 7–3 (18–3). — 2015. — С. 69–72.

36. Торопова, А.В. Анализ согласованности данных в расширенной модели оценки социально-значимого поведения / А.В. Торопова // Региональная информатика (РИ-2016). Юбилейная XV Санкт-Петербургская международная конференция «Региональная информатика (РИ-2016)». (Санкт-Петербург, 26-28 октября 2016 г.): Материалы конференции. — СПб: СПОИСУ. — 2016. — С. 522.

37. Торопова, А.В. Аппарат диагностики согласованности данных в модели социально-значимого поведения / А.В. Торопова // Гибридные и синергетические интеллектуальные системы, Материалы III Всероссийской Поспеловской конференции с международным участием. — 6–11 июня 2016. — Светлогорск. — С. 441–447.

38. Торопова, А.В Диагностика согласованности данных в модели рискованного поведения / А.В. Торопова // Материалы 6-й всероссийской научной конференции по проблемам информатики СПИСОК-2016. (26–29 апреля 2016 г. Санкт-Петербург). — СПб.: ВВМ. — 2016. — С. 566–573.

39. Торопова, А.В. Диагностика согласованности данных респондентов в модели социально-значимого поведения / А.В. Торопова // Материалы 9-й конференции «Информационные технологии в управлении» (ИТУ-2016). — СПб.: АО «Концерн «ЦНИИ «Электроприбор» . — 2016. — С. 620–623.

40. Торопова, А.В. Использование модели социально-значимого поведения со скрытыми переменными в социокомпьютинге / А.В. Торопова // Региональная информатика (РИ-2018). XVI Санкт-Петербургская международная конференция «Региональная информатика (РИ-2018)». (Санкт-Петербург, 24-26 октября 2018 г.): Материалы конференции. — СПб: СПОИСУ. — 2018. — С. 551–552.

41. Торопова, А.В. Использование скрытых переменных в модели социально-значимого поведения / А.В. Торопова // Нечеткие системы, мягкие вычисления и интеллектуальные технология (НСМВИТ-2017): труды VII всероссийской научной-практической конференции (г. Санкт-Петербург, 3–7 июля, 2017 г.). — в 2 т. — Т. 2. — СПб.: Политехника-сервис. — 2017. — С. 159-165.

42. Торопова, А. В. Модель социально-значимого поведения со скрытыми переменными в управлении людскими ресурсами / А.В. Торопова // Материалы конференции «Информационные технологии в управлении» (ИТУ-2018). — СПб.: АО «Концерн «ЦНИИ «Электроприбор». — 2018. — С. 285–289.

43. Торопова, А.В. Подходы к диагностике согласованности данных в байесовских сетях доверия / А.В. Торопова // Труды СПИИРАН. — 2015. — № 6(43) . — С. 156–178.

44. Торопова, А.В. Подходы к диагностике согласованности исходных данных в модели социально-значимого поведения / А.В. Торопова // Материалы 7-й всероссийской научной конференции по проблемам информатики СПИСОК-2017. (26–28 апреля 2017 г. Санкт-Петербург). — СПб.: ВВМ. — 2017. — С. 444–449.

45. Торопова А.В. Сбор данных о последних эпизодах и интенсивности постинга в социальной сети ВКонтакте / А.В. Торопова // Региональная информатика и информационная безопасность. Сборник трудов. Выпуск 9 — СПОИСУ. — СПб., 2020. — ISBN 978-5-907223-89-9. — С. 228–230.

46. Торопова, А.В. Синтез модели социально-значимого поведения как байесовской сети доверия со скрытыми переменными / А.В. Торопова // Информационная безопасность регионов России (ИБРР-2017). X Санкт-Петербургская межрегиональная конференция. (Санкт-Петербург, 1–3 ноября 2017 г.): Материалы конференции. — СПб: СПОИСУ. — 2017. С. 433.

47. Торопова, А.В. Скрытые переменные в модели социально-значимого поведения / А.В. Торопова // Информационная безопасность регионов России (ИБРР-2019). XI Санкт-Петербургская межрегиональная конференция. (Санкт-Петербург, 23–25 октября 2019 г.): Материалы конференции. — СПб.: СПОИСУ. — 2019. — С. 447–448.

48. Торопова, А.В. Машинное обучение байесовской сети доверия как инструмента оценки интенсивности процесса по данным из социальной сети / А.В. Торопова, М.В. Абрамов, Т.В. Тулупьева // 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.

49. Торопова, А.В. Анализ модели социально-значимого поведения со скрытыми переменными / А.В. Торопова, А.В. Суворова // XX Международная конференция по мягким вычислениям и измерениям (SCM-2017). Сборник докладов в 2-х томах. Санкт-Петербург. — 24–26 мая 2017 г. — Т.1. — С. 81–84.

50. Торопова, А.В. Выявление несогласованных данных при оценивании интенсивности социально-значимого поведения / А.В. Торопова, А.В. Суворова // Интеллектуальные системы и технологии: современное состояние и перспективы. Сборник научных трудов III-ей Международной летней школы-семинара по искусственному интеллекту

для студентов, аспирантов и молодых ученых (Тверь-Протасово, 1–5 июля 2015 г.). — Тверь: Изд-во ТвГТУ. — 2015. — С. 119–126.

51. Торопова, А.В. Диагностика входных данных в байесовской сети доверия для оценки параметров социальной активности / А.В. Торопова, А.В. Суворова // Научная сессия НИЯУ МИФИ-2015. Аннотации докладов. В 3 т. — Т.3. — М.: НИЯУ МИФИ. — 2015. — С. 150.

52. Торопова, А.В. Диагностика согласованности входных данных в модели оценивания интенсивности социально-активного поведения / А.В. Торопова, А.В. Суворова // Интегрированные модели и мягкие вычисления в искусственном интеллекте. Сборник научных трудов VIII-й Международной научно-технической конференции (Коломна, 18–20 мая 2015 г.). — М.: Физматлит. — Т.2. — С. 806–815.

53. Торопова, А.В. Диагностика согласованности данных в модели социально-значимого поведения / А.В. Торопова, А.В. Суворова // XIX Международная конференция по мягким вычислениям и измерениям (SCM-2016). Сборник докладов в 2-х томах. — Санкт-Петербург. — 25-27 мая 2016 г. — Т.1. — С. 67–70.

54. Торопова, А.В. Подходы к обработке зашумленных данных в модели социально-значимого поведения / А.В. Торопова, А.В. Суворова // Сборник докладов Международной конференции по мягким вычислениям и измерениям (SCM-2018). Санкт-Петербург. Том 1-2. — Т. 1. — 2018. — С. 138–140.

55. Торопова, А.В. Диагностика согласованности в модели для оценивания интенсивности социально-значимого поведения / А.В. Торопова, А.В. Суворова, А.Л. Тулупьев // Нечеткие системы и мягкие вычисления. — 2015. — Т. 10. — № 1. — С. 93–107.

56. Торопова, А.В. Оценка согласованности данных в модели рискованного поведения / А.В. Торопова, А.В. Суворова, Т.В. Тулупьева // Сборник докладов. XVIII Международная конференция по мягким вычислениям и измерениям SCM-2015 (Санкт-Петербург, 19-21 мая 2015 г.). — СПб.: Издательство СПбГТЭУ «ЛЭТИ» . — 2015. — Том 1. — С. 5–8.

57. Торопова, А.В. Апробация модели интенсивности поведения со скрытыми переменными на данных респондентов о последних публикациях в сети Instagram / А.В. Торопова, Т.В. Тулупьева // XXIV Международная конференция по мягким вычислениям и измерениям (SCM-2021). Сборник докладов. — Санкт-Петербург. — 26–28 мая 2021 г. — СПб.: СПбГЭТУ «ЛЭТИ» . — С. 51–53.

58. Торопова, А.В. Байесовская сеть доверия как модель оценки интенсивности поведения на примере постинга в социальной сети / А.В. Торопова, Т.В. Тулупьева // XXIII Международная конференция по мягким вычислениям и измерениям (SCM-2020). Сборник докладов. — Санкт-Петербург. — 27–29 мая 2020 г. — СПб.: СПбГЭТУ «ЛЭТИ» . — С. 20–22.

59. Торопова, А.В. Дискретизация непрерывной величины, характеризующей интенсивность, в модели социально-значимого поведения / А.В. Торопова, Т.В. Тулупьева // XXV Международная конференция по мягким вычислениям и измерениям (SCM-2022). Сборник докладов. — Санкт-Петербург. — 25–27 мая 2022 г. — СПб.: СПбГЭТУ «ЛЭТИ» . — С. 41–44.

60. Торопова, А.В. Модели оценки интенсивности поведения на примере постинга в социальной сети / А.В. Торопова, Т.В. Тулупьева // VIII Международная научно-практическая конференция «Нечеткие системы, мягкие вычисления и интеллектуальные технологии» НСМВИТ-2020 (29 июня – 1 июля 2020 г., г. Смоленск, Россия). Труды конференции. В 2-х томах. — Т 2. — Смоленск: Универсум. — 2020. — С. 164–172.

61. Торопова, А.В. Апробация модели интенсивности поведения со скрытыми переменными на данных респондентов о последних публикациях в сети Instagram / А.В. Торопова, Т.В. Тулупьева // XXIV Международная конференция по мягким вычислениям и измерениям (SCM-2021). Сборник докладов. — Санкт-Петербург. — 26–28 мая 2021 г. — СПб.: СПбГЭТУ «ЛЭТИ» . — С. 51–53.

62. Торопова, А.В. Программа для диагностики согласованности исходных данных в модели социально-значимого поведения (Input Data Coherence Diagnostics in Behavior Model, Version 01 (IDCDiBM v.01)) / А.В. Торопова, Р.Р. Хайбуллин, А.В. Суворова, А.Л. Тулупьев. — Свидетельство о гос. Регистрации пр. для ЭВМ № 2018615722. — 2018.

63. Тулупьев, А.Л. Алгебраические байесовские сети: система операций локального логико-вероятностного вывода / А.Л. Тулупьев // Информационно-измерительные и управляющие системы. — 2009. — №4. — С. 41–44.

64. Тулупьев, А.Л. Алгебраические байесовские сети: глобальный логико-вероятностный вывод в деревьях смежности: Учеб. Пособие / А.Л. Тулупьев. — СПб.: СПбГУ. — ООО Издательство «Анатолия». — 2007. — 40 с. (Сер. Элементы мягких вычислений).

65. Тулупьев, А.Л. Алгебраические байесовские сети: локальный логико-вероятностный вывод: Учеб. Пособие / А.Л. Тулупьев. — СПб.: СПбГУ. — ООО Издательство «Анатолия». — 2007. — 80 с. (Сер. Элементы мягких вычислений).

66. Тулупьев, А.Л. Алгебраические байесовские сети: система операций глобального логико-вероятностного вывода / А.Л. Тулупьев // Информационно-измерительные и управляющие системы. — 2010. — №11. — С. 65–72.

67. Тулупьев, А.Л. Байесовские сети доверия: логико-вероятностный вывод в ациклических направленных графах / А.Л. Тулупьев, А.В. Сироткин, С.И. Николенко. — СПб.: Изд-во С.-Петерб. ун-та. — 2009. — 400 с.

68. Тулупьев, А.Л. Байесовские сети: логико-вероятностный подход / А.Л. Тулупьев, С.И. Николенко, А.В. Сироткин. — СПб.: Наука. — 2006. — 607 с.

69. Тулупьев, А.Л. Основы теории байесовских сетей: учебник / А.Л. Тулупьев, С.И. Николенко, А.В. Сироткин. — СПб.:Изд-во С.Петерб. ун-та. — 2019. — 399 с.

70. Тулупьев, А.Л. Мягкие вычисления и измерения. Модели и методы: монография / А.Л. Тулупьев, Т.В. Тулупьева, А.В. Суворова, М.В. Абрамов, А.А. Золотин, М.А. Зотов, А.А. Азаров, Е.А. Мальчевская, Д.Г. Левенец, А.В. Торопова, Н.А. Харитонов, А.И. Бирилло, Р.И. Сольницев, С.В. Микони, С.П. Орлов, А.В. Толстов; под ред. д.т.н., проф. С.В. Прокопчиной. — М.: ИД «Научная библиотека», 2017. — 3 т. — 300 с.

71. Тулупьева, Т.В. Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей / Т.В. Тулупьева, А.Е. Пащенко, А.Л. Тулупьев, Т.В. Красносельских, О.С. Казакова. — СПб.: Наука, 2008. — 140 с.

72. Тулупьева, Т.В. Оценка интенсивности поведения респондента в условиях информационного дефицита / Т.В. Тулупьева, А.Л. Тулупьев, А.Е. Пащенко // Труды СПИИРАН. — Вып. 7. — СПб.: Наука. — 2008. — С. 239–254.

73. Фильченков, А.А. Алгебраическая байесовская сеть как основа для медицинской диагностической модели / А.А. Фильченков // «Математическое и компьютерное моделирование в биологии и химии. Перспективы развития». Сборник трудов I Международной интернет-конференции. — Казань.: Из-во «Казанский университет» . — 2012. — С. 162–166.

74. Храмов, М.Ю. Разработка модели поведения пользователей интерактивного интернет-сервиса / М.Ю. Храмов // Экономика и предпринимательство. — 2015. — № 6–1 (59–1). — С. 993–996.

75. Частиков, А.П. Использование байесовской сети при разработке экспертных систем с нечеткими знаниями / А.П. Частиков, И.Ю. Леднева // Информационные технологии в образовании X юбилейная конференция-выставка. Мин-во обр. России, Институт ЮНЕСКО по информационным технологиям в образовании, Институт проблем информатики РАН, Московский комитет обр-ния, Финансовая академия при Правительстве России, Московский госуд. инженерно-физический институт (тех. ун-т). — НПП «БИТ про». — 2000. — С. 359–360.

76. Шевцова, Ю.В. Байесовские технологии в управлении операционными рисками / Ю.В. Шевцова // Электросвязь. — 2010. — № 10. — С. 58–61.

77. Янников, И.М. Оценка экологической ситуации с применением методов математического моделирования / И.М. Янников, М.В. Телегина, Т.Г. Габричидзе // Вектор науки Тольяттинского государственного университета. — 2011. — № 4. — С. 38–41.

78. Ярушкина, Н.Г. Интеллектуальный анализ временных рядов: Учебное пособие / Н.Г. Ярушкина, Т.В. Афанасьева, И.Г. Перфильева. — Ульяновск: УлГТУ. — 2010. — 320 с.

79. Abramov M.V. Social engineering attack modeling with the use of Bayes-ian networks / M.V. Abramov, A.A. Azarov // XIX IEEE International Con-ference on Soft Computing and Measurements (SCM'2016). — St. Peters-burg, 2016. — P. 58–60.

80. Abu-Hanna, A. Prognostic Models in Medicine: AI and Statistical Approaches / A. Abu-Hanna, P.J.F. Lucas // Methods of Information in Medicine. — 2001. — № 40. — P. 1–5.

81. Acid, S. A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service / S. Acid, L.M. de Campos, J.M. Fernández-Luna, S. Rodríguez, J.M. Rodríguez, J. Salcedo // Artificial Intelligence in Medicine. — 2004. — Vol. 30. — №3. — P. 215–232.

82. Acid, S. A hybrid methodology for learning belief networks: BENEDICT / S. Acid, L.M. de Campos // International Journal of Approximate Reasoning. — 2001. — 27. — P. 235–262.

83. Agenarisk [Электронный ресурс]. — Режим доступа: http://www.agenarisk.com/.

84. Akaike, H. A new look at the statistical model identification / H. Akaike // IEEE Transactions on Automatic Control, AC-19. —1974. — P. 716–723. — doi:10.1109/TAC.1974.1100705.

85. Antal, P. Bayesian Networks in Ovarian Cancer Diagnosis: Potentials and Limitations / P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, I. Vergote // IEEE Symposium on Computer-Based Medical Systems, CBMS 2000, proceedings 13th. — 2000. — P. 103–108.

86. Badr, H.S. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study / H.S. Badr, H. Du, M. Marshall, E. Dong, M.M. Squire, L.M. Gardner // Lancet Infectious Diseases. — 2020. — 20(11). — P. 1247–1254. — Doi: 10.1016/S1473-3099(20)30553-3.

87. Baker, C.E. Fathers' and mothers' home literacy involvement and children's cognitive and social emotional development: Implications for family literacy programs / C.E. Baker // Applied Developmental Science. —2013. — 17 (4). — P. 184–197. — Doi:10.1080/10888691.2013.836034.

88. Barabási, A.L. The origin of bursts and heavy tails in human dynamics / A.L. Barabási // Nature. — 2005. — 435. — P. 207–211. — Doi: 10.1038/nature03459.

89. Bari, A. COVID-19 early-alert signals using human behavior alternative data / Bari, A., Khubchandani, A., Wang, J., Heymann, M., Coffee, M. // Social Network Analysis and Mining. —2021. —11, 18. — Doi:10.1007/s13278-021-00723-5.

90. Basturk, B. An Artificial Bee Colony (ABC) Algoritmh for Numeric Function Optimization / B. Basturk, D. Karaboga // IEEE Swarm Intelligence Symposium. — 2006. — P. 10–15.

91. BayesBuilder [Электронный ресурс]. — Режим доступа: http://www.snn.ru.nl/nijmegen/index.php?option=com_content&view=article&id=89&Itemid=212.

92. BayesFusion [Электронный ресурс]. — Режим доступа: https://www.bayesfusion.com/.

93. BayesiaLab [Электронный ресурс]. — Режим доступа:http://www.bayesia.com/.

94. Bayes-Scala [Электронный ресурс]. — Режим доступа: https://github.com/danielkorzekwa/bayes-scala.

95. Bayes Server [Электронный ресурс]. — Режим доступа: http://www.bayesserver.com/.

96. Bayraktarli, Y.Y. On the application of Bayesian probabilistic networks for earthquake risk management [Электронный ресурс] / Y.Y. Bayraktarli, J. Ulfkjaer, U. Yazgan, M.H. Faber // 9th international conference on structural safety and reliability. — Italy, Rome. — Режим доступа: http://www.merci.ethz.ch/Publications/bayota.pdf.

97. BNL [Электронный ресурс]. — Режим доступа: http://www.downscripts.com/bnl_matlab-script.html.

98. bnlearn [Электронный ресурс]. — Режим доступа: https://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf.

99. BNT [Электронный ресурс]. — Режим доступа:https://code.google.com/p/bnt/.

100. Bolger, N. Diary Methods: Capturing Life as it is Lived / N. Bolger, A. Davis, E. Rafaeli // Annu. Rev. Psychol. — 2003. — Vol. 54. — P. 579–616.

101. Bonafede, C.E. Bayesian networks for enterprise risk assessment / C.E. Bonafede, P. Giudici // Physica A: Statistical Mechanics and its Applications. — 2007. — Vol. 382, issue 1. P. 22–28. — Doi: 10.1016/j.physa.2007.02.065.

102. Borsuk, M.E. A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis / M.E. Borsuk, C.A. Stow, K.H. Reckhow // Ecological Modelling. — 2004. — Vol. 173. — No. 2. — P. 219–239.

103. Bouckaert, R.R. Bayesian Belief Networks: from Construction to Inference/ R.R. Bouckaert // PhD thesis. — Utrecht University. — 1995.

104. BUGS [Электронный ресурс]. // MRC Biostatistics Unit. — Режим доступа:http://www.mrc-bsu.cam.ac.uk/software/bugs/.

105. Burnside, E.S. Bayesian Network to Predict Breast Cancer Risk of Mammographic Microcalcifications and Reduce Number of Benign Biopsy Results: Initial Experience / E.S. Burnside, D.L. Rubin, J.P. Fine, R.D. Shachter, G.A. Sisney, W.K. Leung // Radiology. — 2006. — Vol. 240. — № 3. — P. 666–673.

106. Cano-Lozano, M.C. Child-to-Parent Violence: Examining the Frequency and Reasons in Spanish Youth / M.C. Cano-Lozano, S.P. León, L. Contreras // Family Relations. — 2021. — Doi: 10.1111/fare.12567.

107. Carvalho, A.M. Scoring functions for learning Bayesian networks/ A.M. Carvalho. — 2009.

108. Case, K.K. Understanding the modes of transmission model of new HIV infection and its use in prevention planning / K.K. Case, P.D. Ghys, E. Gouws, J.W. Eaton, A. Borquez, J. Stover, P. Cuchi, L.J. Abu-Raddad, G.P. Garnett, T.B. Hallett // Bulletin of the World Health Organization. — 2012. — Vol. 90. — No. 11. — P. 831–838.

109. catnet [Электронный ресурс]. — Режим доступа:https://cran.r-project.org/web/packages/catnet/catnet.pdf.

110. Cebeci, Z. Unsupervised discretization of continuous variables in a chicken egg quality traits dataset / Z. Cebeci, F. Yildiz // Turkish Journal of Agriculture. Food Science and Technology. — 2017. — 5(4). — P. 315–320. — DOI: 10.24925/turjaf.v5i4.315-320.1056.

111. Chickering, D.M. A Transformational Characterization of Equivalent Bayesian Network Structures / D.M. Chickering // UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence. —1995. — P. 87–98.

112. Chickering, D.M. Learning Equivalence Classes of Bayesian Network Structures / D.M. Chickering // Journal of Machine Learning Research. — 2. —445–498. — 2002.

113. Chien, C.-F. Using Bayesian network for fault location on distribution feeder / C.-F. Chien, S.-L. Chen, Y.-S. Lin // IEEE Transactions on Power Delivery. — 17. —785–793. — 2002.

114. Claeskens, G. Model Selection and Model Averaging / G. Claeskens, N. Hjort // Cambridge Series in Statistical and Probabilistic Mathematics. — Cambridge: Cambridge University Press. — 2008. — doi:10.1017/CBO9780511790485.

115. Colombo, D. Order-Independent Constraint-Based Causal Structure Learning / D. Colombo, H.M. Marloes // Journal of Machine Learning Research. — 2014. — vol. 15. — P. 3921-3962.

116. Cooper, G. F. A Bayesian method for the induction of probabilistic networks from data / G. F. Cooper, E. H. Herskovits // Machine Learning. — 9. —1992. — P. 309-347.

117. Cowell, R.G. Modeling operational risk with Bayesian networks / R.G. Cowell, R.J. Verrall, Y.K. Yoon // Journal of Risk and Insurance. — 2007. — Vol. 74. — No 4. — P. 795–827.

118. CRAN Task View: Bayesian Inference [Электронный ресурс]. — Режим доступа:https://cran.r-project.org/web/views/Bayesian.html.

119. Cruyff, M. Accounting for Self-Protective Responses in Randomized Response Data from a Social Security Survey Using the Zero-Inflated Poisson Model / M. Cruyff, U. Bockenholt, A. van den Hout, P. van der Heijden // The Annals of Applied Statistics. — 2008. — Vol. 2. — No. 1. — P. 316–331.

120. Cutler, L. Fathers' Shared Book Reading Experiences: Common Behaviors, Frequency, Predictive Factors, and Developmental Outcomes / L. Cutler. R. Palkovitz // Marriage & Family Review. — 2020. — 56. — 2. — P. 144–173. — DOI: 10.1080/01494929.2019.1683119.

121. Dash D. A Hybrid Anytime Algorithm for the Construction of Causal Models From Sparse Data / D. Dash, M.J. Druzdzel // Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence / K.B. Laskey, H. Prade, editors. —1999. — P. 142–149.

122. deal [Электронный ресурс]. — Режим доступа:https://cran.r-project.org/web/packages/deal/deal.pdf.

123. Dimple [Электронный ресурс]. — Режим доступа:https://github.com/AnalogDevicesLyricLabs/dimple.

124. Dorigo, M. The ant system: Optimization by colony of cooperating agents / M. Dorigo // IEEE Transactions on Systems, Man and Cybernetics. — 1996. — Vol. 26. — no. 1. — P. 1–13.

125. Dorigo, M. Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem / M. Dorigo, L.M. Gambardella // IEEE Transactions on Evolutionary Computation. — 1997. — Vol. 1. — P. 53–66.

126. Eberhart, R.C. A new optimizer using particle swarm theory / R.C. Eberhart, J. Kennedy // Proceedings of the Sixth International Symposium on Micro Machine and Human Science. — 1995. — P. 39–43.

127. Eleye-Datubo, A.G. Marine and Offshore Safety Assessment by Incorporative Risk Modeling in a Fuzzy-Bayesian Network of an Induced Mass Assignment Paradigm / A.G. Eleye-Datubo, A. Wall , J. Wang // Risk Analysis. — 2008. — Vol. 28. — No. 1. — P. 95–112.

128. Elvira [Электронный ресурс]. — Режим доступа: http://leo.ugr.es/elvira/.

129. Factory [Электронный ресурс]. — Режим доступа: http://factorie.cs.umass.edu/.

130. Fan, C.F. BBN-based software project risk management / C.F. Fan, Y.C. Yu // Journal of Systems and Software. — 2004. — Vol. 73. — No 2. — P. 193–203.

131. Feldman, A. Model-Based Diagnostic Decision-Support System for Satellites / A. Feldman, H.V. de Castro, A. van Gemund, G. Provan // Proceedings of The 24th International Workshop on Principles of Diagnosis. — Jerusalem. — October 1-4, 2013. — P. 111–122.

132. Fenton, N. Predicting software defects in varying development lifecycles using Bayesian nets / N. Fenton, M. Neil, W. Marsh, P. Hearty, D. Marquez, P. Krause, R. Mishra //Information and Software Technology. — 2007. — Т. 49. — №. 1. — С. 32–43.

133. Francis, R.A., Bayesian belief networks for predicting drinking water distribution system pipe breaks / R.A. Francis, S.D. Guikema, L. Henneman // Reliability Engineering & System Safety. — 2014. — V. 130. — P. 1 – 11.

134. Friman, P.C. Cooper, Heron, and Heward's applied behavior analysis (2nd ed.): checkered flag for students and professors, yellow flag for the field / P.C. Friman // Applied Behavior Analysis. — 2013. — No. 1. — P. 161–174. — DOI: 10.1901/jaba.2010.43-161.

135. Frowler, F.J. Improving survey questions: design and evaluation / F.J. Frowler // Applied social research methods series. — Thousand Oaks. — CA: SAGE Publications. — 1995. —V. 38. — 200 p.

136. Furlotte, N.A. Quantifying the uncertainty in heritability / N.A. Furlotte, D. Heckerman, C. Lippert // J. Hum. Genet. — 59(5). — 2014. — P. 269–275.

137. Garcia, D. A brief measure to predict exercise behavior: the Archer-Garcia ratio / D. Garcia, T. Daniele, T. Archer // Heliyon. — 2017. — Doi: 10.1016/j.heliyon.2017.e00314.

138. García, S. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning / S. García, J. Luengo, J.A. Sáez, V. López, F. Herrera // IEEE Transactions on Knowledge and Data Engineering. — 2013. — 25. — Doi:10.1109/TKDE.2012.35.

139. Garmaise, M.J. Consumer Default, Credit Reporting, and Borrowing Constraints / M.J. Garmaise, G. Natividad // Journal of Finance. — 2017. —72(5). — P. 2331–2368. — Doi: 10.1111/jofi.12522.

140. Geiger, D. Learning Gaussian Networks / D. Geiger, D. Heckerman // Technical report, Microsoft Research. — Redmond, Washington. — Available as Technical Report MSR-TR-94-10. — 1994.

141. Graham, C.A. Recalling Sexual Behavior: A Methodological Analysis of Memory Recall Bias via Interview using the Diary as the Gold Standard / C.A. Graham, J.A. Catania, R. Brand, T. Duong, J.A. Canchola // Journal of Sex Research. — 2003. — No. 4. — P. 325–332. — DOI: 10.1080/00224490209552198.

142. GraphQL [Электронный ресурс]. — Режим доступа: https://graphql.org.

143. gRain [Электронный ресурс]. — Режим доступа: https://cran.r-project.org/web/packages/gRain/vignettes/gRain-intro.pdf.

144. Hacibeyoglu M., Ibrahim M.H. EF_Unique: An Improved Version of Unsupervised Equal Frequency Discretization Method. Arabian Journal for Science and Engineering. — 2018. — 43. — P. 7695–7704. — DOI: 10.1007/s13369-018-3144-z.

145. Hao, C. Learning Bayesian Network Structure from Data / C. Hao // thesis submitted for the degree of MSc in Mathematics. — Institute of Mathematics Eötvös Loránd University. — Budapest, Hungary. — 2018.

146. Heckerman, D. Learning Bayesian networks: The combination of knowledge and statistical data / D. Heckerman, D. Geiger, D. M. Chickering // Machine Learning. — 1995. — 20. — P. 197–243.

147. Herskovits, E.H. Computer-based probabilistic-network construction / E.H. Herskovits // PhD thesis, Medical Information Sciences. — Stanford Univ. — Stanford, Calif. — 1991.

148. Hofbaur, M. On the Role of Model-based Diagnosis in Functional Safety / M. Hofbaur, M. Sachenbacher // Proceedings of the 24th International Workshop on Principles of Diagnosis. — Jerusalem. — October 1–4. — 2013. P. 65–70.

149. Hugin Expert [Электронный ресурс]. — Режим доступа: http://www.hugin.com/.

150. Infer.NET [Электронный ресурс]. — Режим доступа: http://research.microsoft.com/en-us/um/cambridge/projects/infernet.

151. Jansen, R. A bayesian networks approach for predicting protein-protein interactions from genomic data / R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A.d. Emili, M.b. Snyder, J.F.d. Greenblatt, M. Gerstein // Science. — 2003. — 302(5644). — P. 449–453.

152. Jensen, F.V. Bayesian updating in causa1 probabilistic networks by local computation / F.V. Jensen, S.L. Lauritzen, K.G. Olesen // Computational Statistics Quarterly. — 1990. — 4. — P. 269–82.

153. Ji, J. An artificial bee colony algorithm for learning Bayesian networks / J. Ji, H. Wei, C. Liu // Soft Computing. — 2013. — Vol. 17. — P. 983–994.

154. Ji, Z. A Review of Parameter Learning Methods in Bayesian Network / Z. Ji, Q. Xia, G. Meng // Advanced Intelligent Computing Theories and Applications. ICIC 2015. Lecture Notes in Computer Science / Huang DS., Han K. (eds). — 2015. — vol 9227.

155. Kandasamy, I. Indeterminate Likert scale: feedback based on neutrosophy, its distance measures and clustering algorithm / I. Kandasamy, W.B.V. Kandasamy, J.M. Obbineni, F. Smarandache // Soft Computing. —2020. — 24 (10). — P. 7459–7468. — doi: 10.1007/s00500-019-04372-x.

156. Khlobystova, A.O. Soft Estimates for Social Engineering Attack Propagation Probabilities Depending on Interaction Rates Among Instagram Users / A.O. Khlobystova, M.V. Abramov, A.L. Tulupyev // International Symposium on Intelligent and Distributed Computing. — Springer, Cham. — 2019. — P. 272–277.

157. Khlobystova, A.O. Search for the shortest trajectory of a social engineering attack between a pair of users in a graph with transition probabilities / A.O. Khlobystova, M.V. Abramov, A.L. Tulupyev, A.A. Zolotin // Information and Control Systems. — 2018. — no. 6. P. 74–81.

158. Khodakarami, V. Project Scheduling: Improved approach to incorporate uncertainty using Bayesian Networks / V. Khodakarami, N. Fenton, M. Neil // Project Management Journal. — 2007. — Vol. 38. — No 2. — P. 39–49.

159. Khoshhal, K. Probabilistic Social Behavior Analysis by Exploring Body Motion-Based Patterns / K. Khoshhal, U. Nunes, J. Dias // IEEE Transactions on pattern analysis and machine intelligence. — 2016. — V. 38. — N. 8.

160. Ki, P. School adjustment and academic performance: influences of the interaction frequency with mothers versus fathers and the mediating role of parenting behaviours/ P. Ki // Early Child Development and Care. — 2020. — 190. —7. — P. 1123–1135. — DOI: 10.1080/03004430.2018.1518904.

161. Koelle, D. Applications of bayesian belief networks in social network analysis / D. Koelle, J. Pfautz, M. Farry, Z. Cox, G. Catto, J. Campolongo // Proceedings of the 4th Bayesian Modeling Applications Workshop during the 22nd Annual Conference on Uncertainty in Artificial Intelligence. — 2006.

162. Koller, D. Probabilistic Graphical Models. Principles and Techniques / D. Koller, N. Friedman — Cambridge, Massachusetts. — London: MIT Press. — 2009. — 1231 p.

163. Lacave, C. Knowledge Acquisition in PROSTANET – A Bayesian network for diagnosis prostate cancer / C. Lacave, F.J. Diez // Knowledge-Based Intelligent Information and Engineering Systems. Lecture Notes in Computer Science. — 2003. — Vol. 2774. — P. 1345–1350.

164. Landuyt, D. An ecosystem service approach to support integrated pond management: A case study using Bayesian belief networks / D. Landuyt, P. Lemmens, R. D'hondt, S. Broekx, I. Liekens, T. De Bie, S.A.J. Declerck, De L. Meester, P.L.M. Goethals // Highlighting opportunities and risks // Journal of Environmental Management. —2014. — 145. — P. 79–87.

165. Langseth H. Bayesian networks in reliability / H. Langseth, L. Portinale // Reliability Engineering and System Safety. — 2007. — 92(1). — P. 92–108.

166. Lantos, H. Describing associations between child maltreatment frequency and the frequency and timing of subsequent delinquent or criminal behaviors across development: variation by sex, sexual orientation, and race / H. Lantos, A. Wilkinson, H. Winslow et al. // BMC Public Health 19. —2019. — 1306. —DOI: 10.1186/s12889-019-7655-7.

167. Larranaga, P. Decomposing Bayesian Networks: Triangulation of Moral Graph with Genetic Algorithms / P. Larranaga, C. Kuijpers, M. Poza, R.H. Murga // Statistics and Computing. — 1997. — 7. — P. 19–34. — DOI: 10.1023/A:1018553211613.

168. Lauritzen, S.L. The Em algorithm for graphical association models with missing data / S.L. Lauritzen // Comput. Stat. Data An. —1995. — 19(2). — P. 191–201.

169. Lee, E. Large engineering project risk management using a Bayesian belief network / E. Lee, Y. Park, J.G. Shin // Expert Systems with Applications. — 2009. — Vol. 36. — No 3. — P. 5880–5887.

170. Léger, A. Modeling of human and organizational impacts for system risk analyses / A. Léger, C. Duval, R. Farret, P. Weber, E. Levrat, B. Iung // 9th International Probabilistic Safety Assessment and Management Conference. — Hong Kong, China. — 2008.

171. Lewis, R.W. A semantically constrained Bayesian network for manufacturing diagnosis / R.W. Lewis, R.S. Ransing // International Journal of Production Research. — 1997. — 35:8. — 2171. — 20002188. — DOI: 10.1080/002075497194796.

172. Li, P.C. A fuzzy Bayesian network approach to improve the quantification of organizational influences in HRA frameworks / P.C. Li, G.H. Chen, L.C. Dai, L. Zhang // Safety Science. — 2012. — 50(7). — P. 1569–1583.

173. Liamputtong, P. Handbook of Research Methods in Health Social Sciences /P. Liamputtong. — 2019. — DOI: 10.1007/978-981-10-5251-4.

174. libDAI [Электронный ресурс]. — Режим доступа: https://staff.fnwi.uva.nl/j.m.mooij/libdai/.

175. Xiaotong, L. Bayesian Network Learning and Applications in Bioinformatics / L. Xiaotong // PhD thesis. — University of Kansas. — 2012.

176. Liu, H. A new hybrid method for learning bayesian networks: Separation and reunion / H. Liu, S. Zhou // Knowledge-Based Systems. — 2017. — 121. — P. 185–197.

177. Lucas, P.J.F. Bayesian model-based diagnosis / P.J.F. Lucas // International Journal of Approximate Reasoning. — 27. — 2001. — P. 99–119.

178. Margaritis, D. Learning Bayesian Network Model Structure from Data / D. Margaritis // PhD thesis. — Carnegie Mellon University. — 2003.

179. Mayer, G.R. Behavior analysis for lasting change/ G.R. Mayer, B. Sulzer-Azaroff, M. Wallace. — Cornwall-on-Hudson. — NY: Sloan Publishing. — 2018.

180. McCann, R.K. Bayesian belief networks: applications in ecology and natural resource management / R.K. McCann, B.G. Marcot, R. Ellis // Canadian Journal of Forest Research. — 2006. — Vol. 36. — No 12. — P. 3053–3062.

181. Mengshoel, O.J. Probabilistic model-based diagnosis: An electrical power system case study / O.J. Mengshoel, M. Chavira, K. Cascio, S. Poll, A. Darwiche, S. Uckun // IEEE Trans. on Systems, Man, and Cybernetics. — 09/2010. — 40. — P. 874–885. — DOI: 10.1109/TSMCA.2010.2052037.

182. Mihoub, A. Social behavior modeling based on incremental discrete hidden Markov models / A. Mihoub, G. Bailly, C. Wolf // Human Behavior Understanding, Lecture Notes in Computer Science. — 2013. — P. 172–183.

183. Molloy, B. Developing Bayesian Belief Networks to Support Risk-Based Decision Making in Railway Operations / B. Molloy, N. Balfe, E. Lowe // Conference Proceedings: Applied Human Factors and Ergonomics. — 2014. — Krakow, Poland.

184. Murphy, K. Software Packages for Graphical Models [Электронный ресурс] / K. Murphy. — Режим доступа: http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html.

185. Neapolitan, R.E. Learning Bayesian Networks / R.E. Neapolitan. — Pearson Prentice Hall. — 2003. — 674 p.

186. Neil, M. Using Bayesian networks to model expected and unexpected operational losses / M. Neil, N. Fenton, M. Tailor // Risk Analysis. — 2005. — Vol. 25. — No 4. — P. 963–972.

187. Netica [Электронный ресурс]. — Режим доступа: http://www.norsys.com/.

188. Newsome, D. How contextual behavioral scientists measure and report about behavior: A review of JCBS / D. Newsome, K. Newsome, T.C. Fuller, S. Meyer // Journal of Contextual Behavioral Science. — 2019. — 12. — P. 347–354. — DOI: 10.1016/j.jcbs.2018.11.005.

189. Nicholson, A. [Электронный ресурс] / A. Nicholson, K. Korb. — Режим доступа: http://www.csse.monash.edu.au/bai/book1e/appendix_b.pdf.

190. Nieto-García, M. The More the Merrier? Understanding How Travel Frequency Shapes Willingness to Pay / M. Nieto-García, P.A. Muñoz-Gallego, Ó. Gonzalez-Benito // Cornell Hospitality Quarterly. —2020. — 61(4). — P. 401–415. — Doi:10.1177/1938965519899932.

191. Nikovsky, D. Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics / D. Nikovsky // IEEE Transactions on Knowledge and Data Engineering. — 2000. — Vol. 12. — № 4. — P. 509–516.

192. Niu, D.X. Short-term load forecasting using bayesian neural networks learned by Hybrid Monte Carlo algorithm / D.X. Niu, H.F. Shi, D.D. Wu // Appl. Soft Comput. — 2012. — 12(6). — P. 1822–1827.

193. OpenGM2 [Электронный ресурс]. — Режим доступа: http://hci.iwr.uni-heidelberg.de/opengm2/.

194. Oteniya, L. Diagnosis of Dementia and its Pathologies Using Bayesian Belief Networks / L. Oteniya, J. Cowie, R. Coles // Conference: ICEIS 2006. Proceedings of the Eighth International Conference on Enterprise Information Systems: Databases and Information Systems Integration. — Paphos, Cyprus. — May 23–27, 2006.

195. Perl, J. Causality: Models, Reasoning, and Inference / J. Perl. — Cambridge: Cambridge University Press. — 2000. — 400 p.

196. Perl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference / J. Perl. — NY etc.: Morgan Kaufmann Publ. — 1994. — 552 p.

197. PyMC [Электронный ресурс]. — Режим доступа: http://pymc-devs.github.io/pymc/.

198. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing [Электронный ресурс]. — Vienna, Austria. — Режим доступа: http://www.R-project.org.

199. Radlinski, L. A survey of bayesian net models for software development effort prediction / L. Radlinski // International Journal of Software Engineering and Computing. — 2010. — Vol. 2. — No 2. — P. 95–109.

200. Rao Rajesh, P.N. Neural Models of Bayesian Belief Propagation / P.N. Rao Rajesh // The Bayesian Brain: Probabilistic Approaches to Neural Coding/R. Doya, S. Ishii, A. Pouget, R. N. Rao, eds. — Cambridge, MA: MIT Press. — 2006. — P. 235–319.

201. RDotNet [Электронный ресурс]. — Режим доступа: https://cran.r-project.org/web/packages/rDotNet/index.html.

202. Rehfeldt, R.A. Clarifying the nature and purpose of behavioral assessment: A response to Newsome et al. / R.A. Rehfeldt // Journal of Contextual Behavioral Science. — 2019. — Vol. 14. — P. 37–39.

203. Rhodes, W. Using Booking Data to Model Drug User Arrest Rates: A Preliminary to Estimating the Prevalence of Chronic Drug Use / W. Rhodes, R. Kling, P. Johnston // J Quant Criminol. — 2007. — 23. — P. 1–22.

204. Ross T. New developments in uncertainty assessment and uncertainty management / T. Ross, J. Booker, A. Montoya // Expert Systems with Applications. — 2013. — Vol. 40. — P. 964–974.

205. Rothman, K.J. Epidemiology: An Introduction/ K.J. Rothman. — Oxford etc.: Oxford University Press. — 2002. — 223 p.

206. Ritthaler, M. Bayesian Belief Networks for Astronomical Object Recognition and Classification in CTI-II / M. Ritthaler, G. Luger, R. Young // Astronomical Data Analysis Software and Systems XVI.ASP Conference Series. — Vol. 376. — 2007. — P. 413–416.

207. Roe, B.E. The Validity, Time Burden, and User Satisfaction of the FoodImageTM Smartphone App for Food Waste Measurement Versus Diaries: A Randomized Crossover Trial / B.E. Roe, D. Qi, R.A. Beyl, K.E. Neubig, C.K. Martin, J.W. Apolzan// Resources, Conservation and Recycling. — 2020. — 160. — DOI:10.1016/j.resconrec.2020.104858.

208. Romessis, C. Bayesian network approach for gas path fault diagnosis / C. Romessis, K. Mathioudakis // Journal of engineering for gas turbines and power. — 2006. — 128(1). — P. 64–72.

209. Rstudio [Электронный ресурс]. — Режим доступа: https://rstudio.com/.

210. Ruiz, C. Illustration of the K2 Algorithm for Learning Bayes Net Structures / C. Ruiz. — 2019.

211. Schröder T. Modeling Dynamic Identities and Uncertainty in Social Interactions / T. Schröder, J. Hoey, K.B. Rogers // American Sociological Review. — 2016. — 81(4). — P. 828–855. — Doi:10.1177/0003122416650963.

212. Schwarz, G. Estimating the dimension of a model / G. Schwarz // The Annals of Statistics. —1978. — 6. — P. 461–464. — doi:10.1214/aos/1176344136.

213. Scutari, M. Learning Bayesian Networks with the Bnlearn R Package / M. Scutari // arXiv preprint. — arXiv:0908.3817. — 2009.

214. Scutari, M. Understanding Bayesian Networks with Examples in R / M. Scutari // University of Oxford. — 2017.

215. Shabelnikov, A.N. Interpretability of fuzzy temporal models / A.N. Shabelnikov, S.M. Kovalev, A.V. Sukhanov // Advances in Intelligent Systems and Computing. — 2019. — T. 874. — C. 223–234.

216. Shi, S. Should Buyers or Sellers Organize Trade in a Frictional Market? / S. Shi, A. Delacroix // Quarterly Journal of Economics. — 2018. —133(4). —2171–2214. — Doi:10.1093/qje/qjy009.

217. SimilarWeb [Электронный ресурс]. — Режим доступа: https://www.similarweb.com/website/instagram.com.

218. Singh, M. Construction of Bayesian Network Structures From Data: A Brief Survey and an Efficient Algorithm / M. Singh, M. Valtorta // International Journal of Approximate Reasoning. — 1995. — 12. — P. 111–131.

219. Spiegelhalter, D.J. Bayesian Analysis in Expert Systems / D.J. Spiegelhalter, A.P. Dawid, S.L. Lauritzen, R.G. Cowell // Statistical Science. — 1993. — Vol. 8. — No 3. — P. 219–247.

220. Spirtes, P. An algorithm for fast recovery of sparse causal graphs / P. Spirtes, C. Glymour // Soc. Sci. Comput. Rev. —1991. — 9(1). — P. 62–72.

221. Spirtes, P. Causation, Prediction and Search / P. Spirtes, C. Glymour, R. Scheines. — Springer Verlag, Berlin. — 1993.

222. Stamelos, I. On the use of Bayesian belief networks for the prediction of software productivity / I. Stamelos, L. Angelis, P. Dimou, E. Sakellaris // Information and Software Technology. — 2003. — Vol. 45. — No 1. — P. 51–60.

223. Stan Development Team. 2021. Stan Modeling Language Users Guide and Reference Manual [Электронный ресурс]. — Режим доступа: http://mc-stan.org/.

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

225. Straub, D. Natural hazards risk assessment using Bayesian networks / D. Straub //Safety and Reliability of Engineering Systems and Structures. — 2005. — P. 2535–2542.

226. Tang, A. Predicting Change Impact in Architecture Design with Bayesian Belief Networks / A. Tang, Y. Jin, J. Han, A. Nicholson // Conference: Fifth Working IEEE / IFIP Conference

on Software Architecture (WICSA 2005) . — 6–10 November 2005. — Pittsburgh, Pennsylvania, USA. — DOI: 10.1109/WICSA.2005.51.

227. Tang, A. Using Bayesian belief networks for change impact analysis in architecture design / A. Tang, A. Nicholson, Y. Jin, J. Han // Journal of Systems and Software. — 01/2007. — DOI: 10.1016/j.jss.2006.04.004.

228. The Open Group Base Specifications Issue 7, section 4.16 Seconds Since the Epoch. The Open Group [Электронный ресурс]. — Режим доступа: https://pubs.opengroup.org/onlinepubs/9699919799/xrat/V4_xbd_chap04.html.

229. Titterington, D.M. Bayesian methods for neural networks and related models / D.M. Titterington // Stat. Sci. — 2004. — 19 (1). — P. 128–139.

230. Toropova, A. Data Coherence Diagnosis in BBN Risky Behavior Model / A. Toropova // Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry”(IITI’16). — Springer International Publishing. — 2016. — P. 95–102.

231. Toropova, A.V. Analysis of socially significant behavior model with hidden variables / A.V. Toropova, A.V. Suvorova // 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM). — IEEE. — 2017. — P. 50–53.

232. Toropova, A.V. Data coherence diagnosis in socially significant behavior model / A.V. Toropova, A.V. Suvorova //2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM). — IEEE. — 2016. — P. 14–17.

233. Toropova, A.V. Bayesian Belief Network as a Behavior Intensity Rate Model on the Example of Posting in a Social Network / A.V. Toropova, T.V. Tulupyeva // 2020 XXIII IEEE International Conference on Soft Computing and Measurements (SCM). — St. Petersburg, Russia. — 2020. — P. 22–24. — doi: 10.1109/SCM50615.2020.9198795.

234. Toropova, A.V. Learning Behavior Rate Models on Social Network Data / A.V. Toropova, T.V. Tulupyeva // CEUR Workshop Proceedings. Selected Contributions of the "Russian Advances in Artificial Intelligence" Track at RCAI 2020 co-located with 18th Russian Conference on Artificial Intelligence. — Moscow, Russia. — October 10-16, 2020. — Vol. 2648. — P. 200-209.

235. Toropova, A.V. Synthesis and learning of socially significant behavior model with hidden variables / A.V. Toropova, T.V. Tulupyeva // Advances in Intelligent Systems and Computing. — 2019. — T. 875. — C. 76–84.

236. Toropova, A.V. Testing Behavior Rate Models on data from Vk.com Social Network / A.V. Toropova, T.V. Tulupyeva // CEUR Workshop Proceedings. Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 8th International Conference on “Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT 2020)”. — Smolensk, Russia. — July, 2020. — Vol. 2782. — P. 258–263.

237. Toropova, A., Tulupyeva T. (2021) Comparison of Behavior Rate Models Based on Bayesian Belief Network / A.V. Toropova, T.V. Tulupyeva // 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. — doi: 10.1007/978-3-030-65283-8_42.

238. Troldborg, M. Application of Bayesian Belief Networks to quantify and map areas at risk to soil threats: Using soil compaction as an example / M. Troldborg, I. Aalders, W. Towers, P.D. Hallett, B.M. McKenzie, A.G. Bengough, A. Lilly, B.C. Ball, R.L. Hough // Soil and Tillage Research. — 08/2013. — 132. —56–68. — DOI: 10.1016/j.still.2013.05.005.

239. Trucco, P. A Bayesian Belief Network modeling of organizational factors in risk analysis: a case study in maritime transportation / P. Trucco, E. Cagno, F. Ruggeri, O. Grande // Reliability Engineering and System Safety. — 2008. — No 93. — P. 823–834.

240. Tsamardinos, I. The max-min hill-climbing Bayesian network structure learning algorithm / I. Tsamardinos, L.E. Brown, C.F. Aliferis. // Machine Learning. — 2006. — 65. — P. 31–78.

241. Tulupyeva, T.V. Evidence coherence estimation in risky behavior / T.V. Tulupyeva, A.V. Suvorova, A.V. Toropova, // Soft Computing and Measurements (SCM), 2015 XVIII International Conference. — 2015. — IEEE Conference Publications. — P. 27–29. — DOI: 10.1109/SCM.2015.7190401.

242. Verma, T.S. Equivalence and Synthesis of Causal Models / T.S. Verma, J. Pearl // Uncertainty in Artificial Intelligence. — 1991. — 6. — P. 255-268.

243. Verma, T. An algorithm for deciding if a set of observed independencies has a causal explanation/ / T.S. Verma, J. Pearl // Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence. — Morgan Kaufmann. — 1992. — P. 323-330.

244. Vrieze, S.I. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) /S.I. Vrieze // Psychological Methods. — 2012. — 17. — P. 228–243. — DOI:10.1037/a0027127.

245. Weber, P. Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas / P. Weber, G. Medina-Oliva, C. Simon, B. Iung // Engineering Applications of Artificial Intelligence. — 2012. —25(4). — P. 671–682.

246. Weijters, B. Extremity in horizontal and vertical Likert scale format responses. Some evidence on how visual distance between response categories influences extreme responding / B. Weijters, K. Millet, E. Cabooter // International Journal of Research in Marketing. — 2020. — DOI: 10.1016/j.ijresmar.2020.04.002.

247. Weka [Электронный ресурс]. — Режим доступа: http://www.cs.waikato.ac.nz/ml/weka/.

248. WHO guidelines on physical activity and sedentary behaviour: Web Annex. Evidence profiles. — 2020. — ISBN 978-92-4-001511-1. — 535 p.

249. Wolfson, J.A. Gender differences in global estimates of cooking frequency prior to COVID-19 / J.A. Wolfson, Y. Ishikawa, C. Hosokawa, K. Janisch, J. Massa, D.M. Eisenberg // Appetite. — 2021. —V. 161. — 105117. —Doi: 10.1016/j.appet.2021.105117.

250. XBAIES 2.0 [Электронный ресурс]// Robert Cowell's Personal Home Pages: Software]. — Режим доступа: http://www.staff.city.ac.uk/~rgc/software.html.

251. Yaramakala, S. Speculative Markov Blanket Discovery for Optimal Feature Selection / S. Yaramakala, D. Margaritis // ICDM ’05: Proceedings of the Fifth IEEE International Conference on Data Mining. — 2005. — P. 809–812. — IEEE Computer Society.

252. Yongli, Z. Bayesian network-based approach for power system fault diagnosis / Z. Yongli, H. Limin, L. Jinling // IEEE Transactions on Power Delivery. — 2006. — 21. — P. 634–639.

253. Yoon, Y.K. Modeling operational risk in financial institutions using Bayesian Networks / Y.K. Yoon // A dissertation submitted for the award of the degree of Master of Science in actuarial management. — Cass Business School City of London. — 2003.

254. Zhang S. Urban spatial structure and travel patterns: Analysis of workday and holiday travel using inhomogeneous Poisson point process models / S. Zhang, X. Liu, J. Tang, S. Cheng, Y. Wang // Computers, Environment and Urban Systems. Volume 73. 2019. P. 68–84. Doi: 10.1016/j.compenvurbsys.2018.08.005

255. Zhang, W. The Study of Information Dissemination Behavior in Online Social Network with Propagation Delay / W. Zhang, L. Wang, J. Dai et al. // Journal of Scientific Research & Reports. — 2014. — vol. 3. — no. 19. — P. 2486–2500.

СПИСОК ИЛЛЮСТРАТИВНОГО МАТЕРИАЛА

СПИСОК ТАБЛИЦ

ПРИЛОЖЕНИЕ А СВИДЕТЕЛЬСТВО О РЕГИСТРАЦИИ ПРОГРАММЫ ДЛЯ ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2018615722

Программа для диагностики согласованности исходных данных в модели социально-значимого поведения (Input Data Coherence Diagnostics in Behavior Model, Version 01 (IDCDiBM v.01))

Правообладатели: ***Торопова Александра Витальевна (RU), Хайбуллин Руслан Равилевич (RU), Суворова Алёна Владимировна (RU), Тулупьев Александр Львович (RU)***

Авторы: ***Торопова Александра Витальевна (RU), Хайбуллин Руслан Равилевич (RU), Суворова Алёна Владимировна (RU), Тулупьев Александр Львович (RU)***

Заявка № **2018612734**

Дата поступления **22 марта 2018 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***15 мая 2018 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

ПРИЛОЖЕНИЕ Б ПУБЛИКАЦИИ СОИСКАТЕЛЯ ПО ТЕМЕ ДИССЕРТАЦИИ

Монографии:

1. Тулупьев, А.Л. Мягкие вычисления и измерения. Модели и методы: монография / А.Л. Тулупьев, Т.В. Тулупьева, А.В. Суворова, М.В. Абрамов, А.А. Золотин, М.А. Зотов, А.А. Азаров, Е.А. Мальчевская, Д.Г. Левенец, А.В. Торопова, Н.А. Харитонов, А.И. Бирилло, Р.И. Сольницев, С.В. Микони, С.П. Орлов, А.В. Толстов; под ред. д.т.н., проф. С.В. Прокопчиной. — М.: ИД «Научная библиотека», 2017. — 3 т. — 300 с.

Статьи, опубликованные в журналах из перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук:

2. Торопова, А.В. Байесовские сети доверия: инструменты и использование в учебном процессе / А.В. Торопова // Компьютерные инструменты в образовании. — №4. — 2016. — С. 43–53.

3. Торопова, А.В. Подходы к диагностике согласованности данных в байесовских сетях доверия / А.В. Торопова // Труды СПИИРАН. — 2015. — № 6(43) . — С. 156–178.

4. Торопова, А.В. Машинное обучение байесовской сети доверия как инструмента оценки интенсивности процесса по данным из социальной сети / А.В. Торопова, М.В. Абрамов, Т.В. Тулупьева // 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.

5. Столярова, В.Ф. Модель для оценки частоты публикации постов в онлайн социальной сети по неполным данным с учетом объективных детерминант поведения / В.Ф. Столярова, А.В. Торопова, А.Л. Тулупьев // Нечеткие системы и мягкие вычисления. — 2021. — Т. 16. — № 2. — С. 77–95. — Doi: 10.26456/fssc81.

Статьи, опубликованные в изданиях, индексируемых WoS/Scopus:

6. Toropova, A.V. Discretization of a Continuous Frequency Value in a Model of Socially Significant Behavior / A.V. Toropova, T.V. Tulupyeva // 2022 XXV International Conference on Soft Computing and Measurements (SCM). — St. Petersburg, Russia. — 2022. — P. 28–30. — Doi: 10.1109/SCM55405.2022.9794892.

7. Toropova, A., Comparison of Behavior Rate Models Based on Bayesian Belief Network / A.V. Toropova, T.V. Tulupyeva // Dolinina O. et al. (eds) Recent Research in Control Engineering and Decision Making. ICIT 2020. — Studies in Systems, Decision and Control. —2021. — Vol 337. — Springer, Cham. — Doi: 10.1007/978-3-030-65283-8_42.

8. Toropova, A.V. Approbation of the behavior rate model with hidden variables based on respondents' data on recent Instagram posts / A.V. Toropova, T.V. Tulupyeva // 2021 XXIV International Conference on Soft Computing and Measurements (SCM). — St. Petersburg, Russia. — 2021. — P. 43–45. — Doi: 10.1109/SCM52931.2021.9507171.

9. Toropova, A.V. Testing Behavior Rate Models on data from Vk.com Social Network / A.V. Toropova, T.V. Tulupyeva // CEUR Workshop Proceedings. Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 8th International Conference on "Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT 2020)". — Smolensk, Russia. — July, 2020. — Vol. 2782. — P. 258–263.

10. Toropova, A.V. Bayesian Belief Network as a Behavior Intensity Rate Model on the Example of Posting in a Social Network / A.V. Toropova, T.V. Tulupyeva // 2020 XXIII IEEE International Conference on Soft Computing and Measurements (SCM). — St. Petersburg, Russia. — 2020. — P. 22–24. — Doi: 10.1109/SCM50615.2020.9198795.

11. Toropova, A.V. Learning Behavior Rate Models on Social Network Data / A.V. Toropova, T.V. Tulupyeva // CEUR Workshop Proceedings. Selected Contributions of the "Russian Advances in Artificial Intelligence" Track at RCAI 2020 co-located with

18th Russian Conference on Artificial Intelligence. — Moscow, Russia. — October 10-16, 2020. — Vol. 2648. — P. 200–209.

12. Toropova, A.V. Synthesis and learning of socially significant behavior model with hidden variables / A.V. Toropova, T.V. Tulupyeva // Advances in Intelligent Systems and Computing. — 2019. — V. 875. — P. 76–84.

13. Toropova, A. Data Coherence Diagnosis in BBN Risky Behavior Model / A. Toropova // Proceedings of the First International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'16). — Springer International Publishing. — 2016. — P. 95–102.

14. Toropova, A.V. Analysis of socially significant behavior model with hidden variables / A.V. Toropova, A.V. Suvorova // 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM). — IEEE. — 2017. — P. 50–53.

15. Toropova, A.V. Data coherence diagnosis in socially significant behavior model / A.V. Toropova, A.V. Suvorova // 2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM). — IEEE. — 2016. — P. 14–17.

16. Toropova, A.V. Evidence coherence estimation in risky behavior / A.V. Toropova, A.V. Suvorova, T.V. Tulupyeva // Soft Computing and Measurements (SCM), 2015 XVIII International Conference. — 2015. — IEEE Conference Publications. — P. 27–29. — Doi: 10.1109/SCM.2015.7190401.

Научные тезисы и доклады, опубликованные в других изданиях:

17. Торопова, А.В. Анализ согласованности данных в модели оценки интенсивности социально-значимого поведения / А.В. Торопова // Актуальные направления научных исследований XXI века: теория и практика. — Том 3. — 7–3 (18–3). — 2015. — С. 69–72.

18. Торопова, А.В. Анализ согласованности данных в расширенной модели оценки социально-значимого поведения / А.В. Торопова // Региональная информатика (РИ-2016). Юбилейная XV Санкт-Петербургская международная конференция «Региональная информатика (РИ-2016)». (Санкт-Петербург, 26-28 октября 2016 г.): Материалы конференции. — СПб: СПОИСУ. — 2016. — С. 522.

19. Торопова, А.В. Аппарат диагностики согласованности данных в модели социально-значимого поведения / А.В. Торопова // Гибридные и синергетические интеллектуальные системы, Материалы III Всероссийской Поспеловской конференции с международным участием. — 6–11 июня 2016. — Светлогорск. — С. 441–447.

20. Торопова, А.В Диагностика согласованности данных в модели рискованного поведения / А.В. Торопова // Материалы 6-й всероссийской научной конференции по проблемам информатики СПИСОК-2016. (26–29 апреля 2016 г. Санкт-Петербург). — СПб.: ВВМ. — 2016. — С. 566–573.

21. Торопова, А.В. Диагностика согласованности данных респондентов в модели социально-значимого поведения / А.В. Торопова // Материалы 9-й конференции «Информационные технологии в управлении» (ИТУ-2016). — СПб.: АО «Концерн «ЦНИИ «Электроприбор» . — 2016. — С. 620–623.

22. Торопова, А.В. Использование модели социально-значимого поведения со скрытыми переменными в социокомпьютинге / А.В. Торопова // Региональная информатика (РИ-2018). XVI Санкт-Петербургская международная конференция «Региональная информатика (РИ-2018)». (Санкт-Петербург, 24-26 октября 2018 г.): Материалы конференции. — СПб: СПОИСУ. — 2018. — С. 551–552.

23. Торопова, А.В. Использование скрытых переменных в модели социально-значимого поведения / А.В. Торопова // Нечеткие системы, мягкие вычисления и интеллектуальные технология (НСМВИТ-2017): труды VII всероссийской научной-практической конференции (г. Санкт-Петербург, 3–7 июля, 2017 г.). — в 2 т. — Т. 2. — СПб.: Политехника-сервис. — 2017. — С. 159-165.

24. Торопова, А. В. Модель социально-значимого поведения со скрытыми переменными в управлении людскими ресурсами / А.В. Торопова // Материалы конференции «Информационные технологии в управлении» (ИТУ-2018). — СПб.: АО «Концерн «ЦНИИ «Электроприбор». — 2018. — С. 285–289.

25. Торопова, А.В. Подходы к диагностике согласованности исходных данных в модели социально-значимого поведения / А.В. Торопова // Материалы 7-й всероссийской научной конференции по проблемам информатики СПИСОК-2017. (26–28 апреля 2017 г. Санкт-Петербург). — СПб.: ВВМ. — 2017. — С. 444–449.

26. Торопова, А.В. Синтез модели социально-значимого поведения как байесовской сети доверия со скрытыми переменными / А.В. Торопова // Информационная безопасность регионов России (ИБРР-2017). X Санкт-Петербургская межрегиональная конференция. (Санкт-Петербург, 1–3 ноября 2017 г.): Материалы конференции. — СПб: СПОИСУ. — 2017. С. 433.

27. Торопова, А.В. Скрытые переменные в модели социально-значимого поведения / А.В. Торопова // Информационная безопасность регионов России (ИБРР-2019). XI Санкт-Петербургская межрегиональная конференция. (Санкт-Петербург, 23–25 октября 2019 г.): Материалы конференции. — СПб.: СПОИСУ. — 2019. — С. 447–448.

28. Торопова, А.В. Анализ модели социально-значимого поведения со скрытыми переменными / А.В. Торопова, А.В. Суворова // XX Международная конференция по мягким вычислениям и измерениям (SCM-2017). Сборник докладов в 2-х томах. Санкт-Петербург. — 24–26 мая 2017 г. — Т.1. — С. 81–84.

29. Торопова, А.В. Выявление несогласованных данных при оценивании интенсивности социально-значимого поведения / А.В. Торопова, А.В. Суворова // Интеллектуальные системы и технологии: современное состояние и перспективы. Сборник научных трудов III-ей Международной летней школы-семинара по искусственному интеллекту для студентов, аспирантов и молодых ученых (Тверь-Протасово, 1–5 июля 2015 г.). — Тверь: Изд-во ТвГТУ. — 2015. — С. 119–126.

30. Торопова, А.В. Диагностика входных данных в байесовской сети доверия для оценки параметров социальной активности / А.В. Торопова, А.В. Суворова // Научная сессия НИЯУ МИФИ-2015. Аннотации докладов. В 3 т. — Т.3. — М.: НИЯУ МИФИ. — 2015. — С. 150.

31. Торопова, А.В. Диагностика согласованности входных данных в модели оценивания интенсивности социально-активного поведения / А.В. Торопова, А.В. Суворова // Интегрированные модели и мягкие вычисления в искусственном интеллекте. Сборник научных трудов VIII-й Международной научно-технической конференции (Коломна, 18–20 мая 2015 г.). — М.: Физматлит. — Т.2. — С. 806–815.

32. Торопова, А.В. Диагностика согласованности данных в модели социально-значимого поведения / А.В. Торопова, А.В. Суворова // XIX Международная конференция по мягким вычислениям и измерениям (SCM-2016). Сборник докладов в 2-х томах. — Санкт-Петербург. — 25-27 мая 2016 г. — Т.1. — С. 67–70.

33. Торопова, А.В. Подходы к обработке зашумленных данных в модели социально-значимого поведения / А.В. Торопова, А.В. Суворова // Сборник докладов Международной конференции по мягким вычислениям и измерениям (SCM-2018). Санкт-Петербург. Том 1-2. — Т. 1. — 2018. — С. 138–140.

34. Торопова, А.В. Оценка согласованности данных в модели рискованного поведения / А.В. Торопова, А.В. Суворова, Т.В. Тулупьева // Сборник докладов. XVIII Международная конференция по мягким вычислениям и измерениям SCM-2015 (Санкт-Петербург, 19-21 мая 2015 г.). — СПб.: Издательство СПбГТЭУ «ЛЭТИ» . — 2015. — Том 1. — С. 5–8.

35. Торопова, А.В. Байесовская сеть доверия как модель оценки интенсивности поведения на примере постинга в социальной сети / А.В. Торопова, Т.В. Тулупьева // XXIII Международная конференция по мягким вычислениям и измерениям (SCM-2020). Сборник докладов. — Санкт-Петербург. — 27–29 мая 2020 г. — СПб.: СПбГЭТУ «ЛЭТИ» . — С. 20–22.

36. Торопова, А.В. Модели оценки интенсивности поведения на примере постинга в социальной сети / А.В. Торопова, Т.В. Тулупьева // VIII Международная научно-практическая конференция «Нечеткие системы, мягкие вычисления и интеллектуальные технологии» НСМВИТ–2020 (29 июня – 1 июля 2020 г., г.

Смоленск, Россия). Труды конференции. В 2-х томах. — Т 2. — Смоленск: Универсум. — 2020. — С. 164–172.

37. Торопова, А.В. Сбор данных о постинге в социальной сети ВКонтакте для апробации модели интенсивности поведения со «следующим» эпизодом / А.В. Торопова // Региональная информатика (РИ–2020). XVII Санкт-Петербургская международная конференция «Региональная информатика (РИ-2020)». Материалы конференции. Часть 2. — Санкт-Петербург, Россия. — 28–30 октября 2020 г. — СПб: СПОИСУ. — С. 263.

38. Торопова А.В. Сбор данных о последних эпизодах и интенсивности постинга в социальной сети ВКонтакте / А.В. Торопова // Региональная информатика и информационная безопасность. Сборник трудов. Выпуск 9 — СПОИСУ. — СПб., 2020. — ISBN 978-5-907223-89-9. — С. 228–230.

39. Торопова, А.В. Апробация модели интенсивности поведения со скрытыми переменными на данных респондентов о последних публикациях в сети Instagram / А.В. Торопова, Т.В. Тулупьева // XXIV Международная конференция по мягким вычислениям и измерениям (SCM-2021). Сборник докладов. — Санкт-Петербург. — 26–28 мая 2021 г. — СПб.: СПбГЭТУ «ЛЭТИ» . — С. 51–53.

40. Торопова, А.В. Дискретизация непрерывной величины, характеризующей интенсивность, в модели социально-значимого поведения / А.В. Торопова, Т.В. Тулупьева // XXV Международная конференция по мягким вычислениям и измерениям (SCM-2022). Сборник докладов. — Санкт-Петербург. — 25–27 мая 2022 г. — СПб.: СПбГЭТУ «ЛЭТИ» . — С. 41–44.

Зарегистрированные программы для ЭВМ в Роспатент:

41. Торопова, А.В. Программа для диагностики согласованности исходных данных в модели социально-значимого поведения (Input Data Coherence Diagnostics in Behavior Model, Version 01 (IDCDiBM v.01)) / А.В. Торопова, Р.Р. Хайбуллин, А.В. Суворова, А.Л. Тулупьев. — Свидетельство о гос. Регистрации пр. для ЭВМ № 2018615722. — 2018.

Иные публикации:

42. Торопова, А.В. Диагностика согласованности в модели для оценивания интенсивности социально-значимого поведения / А.В. Торопова, А.В. Суворова, А.Л. Тулупьев // Нечеткие системы и мягкие вычисления. — 2015. — Т. 10. — № 1. — С. 93–107.

ПРИЛОЖЕНИЕ В АКТЫ О ВНЕДРЕНИИ

МИНОБРНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ «САНКТ-ПЕТЕРБУРГСКИЙ ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР РОССИЙСКОЙ АКАДЕМИИ НАУК» (СПБ ФИЦ РАН)

14-я линия В.О., д. 39, Санкт-Петербург, 199178
Телефон: (812) 328-33-11, факс: (812) 328-44-50, Email: info@ spcras.ru, https://spcras.ru/
ОКПО:04683303, ОГРН:1027800514411, ИНН/КПП:7801003920/780101001

УТВЕРЖДАЮ
заместитель директора
по научной работе СПб ФИЦ РАН
_____ С.В. Кулешов
«21» 09 2022

А К Т

Об использовании результатов диссертационной работы Тороповой Александры Витальевны «Методы и алгоритмы обработки неопределенности данных при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений» в научно-исследовательской работе СПб ФИЦ РАН

Комиссия в составе: председателя – д.т.н., профессора Осипова В.Ю., членов комиссии: к.т.н. Зайцевой А.А. и к.воен.н. Силлы Е.П. составила настоящий акт о том, что научные результаты, полученные Тороповой Александрой Витальевной в процессе выполнения диссертационной работы «Методы и алгоритмы обработки неопределенности данных при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений», а именно:

- метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанной на байесовской сети доверия;
- алгоритм обработки некорректности информации об эпизодах поведения, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида;
- метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений;
- архитектура и прототип комплекса программ, реализующие разработанные методы и алгоритмы;

были внедрены в научно-исследовательской работе СПб ФИЦ РАН № 0073-2019-0003 «Состояние и перспективы развития информационного общества и цифровой экономики в России», в рамках которой решалась задача автоматизации сбора информации об эпизодическом поведении индивида и вычислении сводных характеристик такого поведения. В условиях ограниченности ресурсов особую роль играют предложенные соискателем методы и алгоритмы учета неопределенности, сопутствующей самоотчетам респондентов. Реализация положений, выносимых на защиту в диссертационной работе, позволила повысить точность получаемых оценок интенсивности пуассоновского процесса по ограниченному объему

доступных наблюдений по сравнению с предыдущими результатами проектов, реализованных в лаборатории теоретических и междисциплинарных проблем информатики СПб ФИЦ РАН. Комиссия отмечает теоретическую и практическую значимость полученных в диссертационной работе научных результатов.

Председатель комиссии:

директор СПИИРАН,
д.т.н., профессор
Осипов Василий Юрьевич

Члены комиссии:

Ученый секретарь СПИИРАН,
к.воен.н
Силла Евгений Петрович

ученый секретарь СПб ФИЦ РАН
к.т.н.
Зайцева Александра Алексеевна

Северо-Западный институт управления — филиал Федерального государственного бюджетного образовательного учреждения высшего образования «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации»

АКТ РЕАЛИЗАЦИИ

результатов диссертационного исследования на соискание ученой степени кандидата технических наук Тороповой Александры Витальевны

Комиссия в составе:

Председатель – декан факультета государственного и муниципального управления СЗИУ РАНХиГС к.э.н. Лихтин А.А.

Члены комиссии:

- Заведующий кафедрой государственного и муниципального управления к.ф.н. Катанандов С.Л.
- Руководитель образовательного направления «Государственное и муниципальное управление» к.полит.н. Тирабян К.К.

Составила настоящий акт о том, что результаты диссертационного исследования Тороповой А.В. **«Методы и алгоритмы обработки неопределенности данных при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений»**, а именно:

- метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанной на байесовской сети доверия;
- алгоритм обработки некорректности информации об эпизодах поведения, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида;

- прототип комплекса программ, реализующий разработанные методы и алгоритмы

используются в учебном процессе факультета государственного и муниципального управления Северо-Западный институт управления, филиала Федерального государственного бюджетного образовательного учреждения высшего образования «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации» по программе второго высшего образования при проведении практических и теоретических занятий по дисциплине «Стратегия управления человеческими ресурсами».

Председатель комиссии

Декан факультета ГМУ Лихтин А.А.

Члены комиссии:

Зав. Кафедрой ГМУ Катанандов С.Л.

Руководитель образовательного направления «ГМУ» Тирабян К.К.

09.09.22

Общество с ограниченной ответственностью «ХОУМ ФИТНЕС»

Акт

Внедрения результатов диссертационной работы Тороповой А.В. на соискание учёной степени кандидата технических наук

Я, Учредитель ООО «ХОУМ ФИТНЕС», Тремпольская Ганна Богдановна, настоящим подтверждаю, что результаты диссертационного исследования «Разработка методов и алгоритмов обработки неопределенности данных при оценке интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений» на соискание учёной степени кандидата технических наук Тороповой А.В., в частности программное обеспечение, реализующее алгоритм обработки некорректности информации об эпизодах поведения, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, использовались при разработке подходящего для клиента режима физических нагрузок.

Учредитель ООО «ХОУМ ФИТНЕС» Тремпольская Г. Б.

14.07.22

г. Санкт-Петербург, вн.тер.г. муниципальный округ Юнтолово, ул Плесецкая, д. 10, стр. 1, помещ. 27-Н

+7 (995) 230-70-70, homefit.taplink.ws

ИНН 7814792358/ КПП 781401001