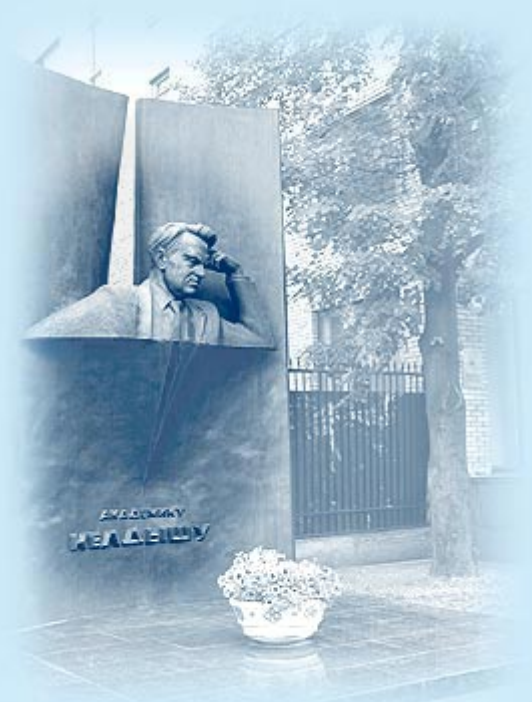




ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 60 за 2022 г.



ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

В.И. Балута

Формирование синтетических
данных при подготовке
обучающих множеств для
систем предупреждения ЧС

Статья доступна по лицензии
[Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)



Рекомендуемая форма библиографической ссылки: Балута В.И. Формирование синтетических данных при подготовке обучающих множеств для систем предупреждения ЧС // Препринты ИПМ им. М.В.Келдыша. 2022. № 60. 28 с. <https://doi.org/10.20948/prepr-2022-60>
<https://library.keldysh.ru/preprint.asp?id=2022-60>

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

В.И.Балута

**Формирование синтетических данных
при подготовке обучающих множеств
для систем предупреждения ЧС**

Москва — 2022

Балута В.И.

Формирование синтетических данных при подготовке обучающих множеств для систем предупреждения ЧС

Настоящий препринт включает краткий обзор информационных материалов по проблеме формирования синтетических данных и обоснование возможности применения математического моделирования для подготовки обучающих множеств в интересах создания средствами машинного обучения прогнозных и аналитических моделей, применяемых в системах предупреждения и реагирования на ЧС.

Дана оценка современного состояния исследований в области машинного обучения, проведен анализ и обобщение практических методов получения синтетических данных для формирования обучающих множеств с детализацией особенностей применительно к представлению информации в виде числовых, текстовых или образных форматов, выработаны рекомендации по использованию различных механизмов создания синтетических данных в целях подготовки обучающих множеств прогнозных и аналитических моделей по основным видам угроз.

Ключевые слова: искусственный интеллект, синтетические данные, генерация данных, машинное обучение, математическое моделирование.

Victor Ivanovich Baluta

Formation of synthetic data in the preparation of training sets for emergency warning systems

This work is devoted to the review and analysis of information materials on the problem of synthetic data formation in the preparation of training sets of predictive and analytical models in emergency response systems.

A brief assessment of the current state of research in the field of machine learning is given, an analysis and generalization of practical methods for obtaining synthetic data for the formation of training sets with detailed features in relation to the presentation of information in various forms is carried out, recommendations are made on the use of various mechanisms for creating synthetic data in order to prepare training sets of predictive and analytical models for the main types of threats based on expert ratings.

Key words: artificial intelligence, synthetic data, data generation, machine learning, mathematical modeling.

Введение

Одним из следствий технологического развития общества является постоянное усложнение производственных, организационных, социальных систем и процессов их взаимодействий, порождающее существенное возрастание неопределённости при оценке и прогнозировании их поведения и затрудняющее своевременную выработку адекватных управленческих решений. Неопределённость при принятии управленческих решений может порождаться невозможностью одновременного охвата и учета всей структуры взаимосвязей управляемого объекта, недостатком знаний о закономерностях его поведения, дефицитом времени на анализ и прогноз развития ситуации на основе имеющихся данных, нехваткой требующейся для выработки решений информации, наличием избыточного зашумления информационных потоков, слабой предсказуемостью траекторий изменения параметров внешней среды, наличием вероятности целенаправленного противодействия со стороны конкурирующих систем и рядом других обстоятельств, которые так или иначе влияют на оценку текущей обстановки и адекватность реагирования. По природе возникновения чрезвычайных ситуаций (далее – ЧС) принято выделять техногенные, природные, экологические, антропогенные, социальные и комбинированные.

В кризисных или чрезвычайных ситуациях одной из ключевых задач системы управления является своевременность выявления предпосылок ЧС, оперативность оценки обстановки, прогноз ее ожидаемого развития и выбор рациональных путей реагирования с целью минимизации вероятных негативных последствий. До сих пор основной метод решения этой задачи – экспертный, то есть опора на опыт и знания специалистов. Однако уровень подготовки и опыта управленческого персонала не всегда адекватен сложности возникающих задач. Кроме того, в силу особенностей мышления человек не способен по отрывочным данным оценить резко меняющуюся обстановку во всей ее многогранности. А потому при анализе ситуации ориентируется на сравнительно небольшое количество показателей, представляющихся наиболее важными или достоверными. Такой подход часто приводит к задержкам в выработке решений или к ошибочным действиям.

В качестве поддерживающей технологии в последние десятилетия во многих сферах предлагается математическое моделирование, позволяющее воспроизводить различные аспекты обстановки в их взаимосвязи. Однако непосредственное применение этого подхода также сопряжено с необходимостью преодоления ряда проблем. Прежде всего, стремление к получению адекватных результатов требует использования моделей высокого уровня сложности, а это зачастую приводит к тому, что сам процесс получения результатов по длительности может превышать реальное время протекания моделируемых процессов, что уже делает применение этого подхода проблемным в конкретных ситуациях. Однако более существенным недостатком прямого использования математических моделей является невозможность их

применения в условиях неопределённости, которые характерны практически для любого случая ЧС. Как правило, каждая ЧС имеет свои уникальные особенности, а информация о причинах, характере и деталях явлений, выходящих за рамки привычного хода событий, является неполной, отрывочной, а зачастую – и противоречивой, что не позволяет однозначно сформировать полный перечень значений исходных данных для математической модели, а неточности на основе допущений при постановке чреваты значительными отклонениями от реальности прогнозных параметров развития ситуации и сопутствующих последствий, что также повышает риск выработки неверных решений в сложной обстановке.

В связи с чем вполне понятно то внимание, которое сейчас уделяется разработке средств искусственного интеллекта (далее - ИИ) для применения в системах кризисного управления. В «Национальной стратегии развития искусственного интеллекта на период до 2030 года» сделан акцент на необходимости интенсивного развития и внедрения средств ИИ в государственном управлении. Системы на основе ИИ ориентированы на решение узкоспециализированных задач, при этом от обычного математического моделирования их отличает как оперативность получения результатов, так и возможность их использования в условиях неполноты или неточности информации. Хорошо обученная нейронная сеть способна не только «оценить обстановку» по совокупности разрозненных признаков на базе имеющейся информации, но и уточнять её по мере увеличения потока поступающих данных. В то же время, в силу закона сохранения сложности более высокая эффективность систем на базе ИИ достигается за счёт больших трудозатрат на этапе создания подобных инструментов. В числе проблемных вопросов находится и формирование обучающих множеств.

В настоящее время машинное обучение (далее – МО) – интенсивно развивающаяся сфера научно-практической деятельности: фактически непрерывно в реальном времени появляются сведения о разработке новых или о совершенствовании апробированных ранее подходов, архитектур, методов, расширяющих возможности решения самых разнообразных задач с применением обученных алгоритмов. Как отмечают специалисты-практики в сфере машинного обучения, до настоящего времени нет четких научно-обоснованных критериев по количественным и качественным параметрам обучающих множеств. Иногда для достижения требуемого уровня обучения приходится увеличивать объем данных на порядок, в других ситуациях бывает достаточно изменить состав или объем выборки на 5-10%, чтобы получить хорошо обученную модель.

Поскольку системы на основе машинного обучения предназначены для решения практических задач, то в основе обучающих множеств лежат данные, полученные в результате натурных наблюдений. Работа по сбору таких данных в необходимом объеме с последующей обработкой в нужном ракурсе для решения конкретной узкоспециализированной задачи является не только весьма

трудоемкой, но в ряде случаев и трудноразрешимой. Например, в интересах решения задач в сфере предупреждения угроз безопасности населения критически недостаточно необходимого объема реальных данных, полученных в натурных условиях. Во-первых, сами по себе ЧС относятся к классу редких событий, что уже ограничивает возможности получения большого количества натурных данных. Во-вторых, случаи ЧС – это не целенаправленные эксперименты, в которых возможно всестороннее исследование происходящих процессов с фиксацией значений множества параметров. Как правило, какие-то данные о причинах, параметрах, характеристиках произошедших при ЧС событий собираются постфактум и представляют собой разрозненную картину отрывочных сведений, извлекаемых из сообщений, рассказов, видеозаписей, фиксации каких-то инструментальных измерений и т.п.

Очевидно, что восполнить их объем путем натурных экспериментов нереально и по этическим, и по экономическим соображениям. Но даже если не учитывать эти соображения, крупномасштабные эксперименты подобного рода бессмысленны ввиду уникальности каждого события ЧС. По большинству вероятных негативных сценариев ЧС можно опираться лишь на умозрительные оценки экспертов. Поэтому использование вместо натурных данных их синтетического аналога на основе экспертных оценок и сценариев представляется единственно возможным способом создания подобных обучающих множеств.

Аргументом в пользу этого подхода служит и постоянно расширяющаяся практика дополнения обучающих множеств искусственными (синтетическими) данными. Судя по многочисленным публикациям, разрабатываются различные технологии их формирования, проводятся оценки эффективности этих технологий применительно к решению типовых задач. Основным источником информации о последних достижениях и практических рекомендациях по большей части являются сообщения на площадках обмена опытом специалистов-практиков в интернет-пространстве.

На основании изучения таких источников можно получить общее представление о путях, способах и технических приёмах разрешения сложностей, возникающих при формировании обучающих множеств. На основании анализа сложившихся подходов и собственного опыта можно рассмотреть возможные пути комплексного и целенаправленного получения исходных данных и формирования на их основе синтетических массивов в интересах создания обучающих множеств прогнозных и аналитических моделей по основным видам угроз в интересах разработки инструментов поддержки принятия решений в ЧС.

Другими словами, речь должна вестись о развитии систем **«гибридного интеллекта»**, органично объединяющего в своей основе знания и опыт экспертов, теоретические расчетные модели и средства машинного обучения.

В качестве примечания: представленные ниже материалы сопровождаются рисунками, которые имеют иллюстративный характер и извлечены либо из цитируемых работ, либо из свободного интернета.

Понятие синтетических данных

В литературе можно встретить несколько определений этого понятия, достаточно близких по своему смысловому содержанию. В наиболее общем виде считается, что любые данные, которые не получены методом прямого измерения, считаются синтетическими [1]. В большинстве определений упор делается на том, что синтетические данные в отличие от натуральных генерируются программными алгоритмами [2, 3]. Прежде всего, к таковым относят сгенерированные специальными программами просто «сходные» с натуральными или имитирующие реальные, но не имеющие отношения к реальности, данные. В этом случае синтетические данные – это новые данные, полученные из какого-то заданного эталонного распределения реальных данных. Синтетическими также считаются любые искусственные данные, полученные с помощью каких-либо способов переработки исходных натуральных или модельных данных. В частности, к таким методам переработки относятся многочисленные способы аугментации (этот термин означает увеличение, размножение, раздутие) данных путем внесения различного рода искажений в исходные данные. В сфере МО выделилось целое направление по созданию обучающих массивов синтетических данных, разработаны платформы для их генерации, применяющие различные технологии вплоть до использования обученных нейронных сетей.

Отмечается, что подобная генерация синтетических данных имеет как свои достоинства, так и недостатки. В частности, ключевым преимуществом является более быстрое и дешевое создание наборов данных практически любого объема; возможность обойти или смягчить проблемы, связанные с предвзятостью, конфиденциальностью и лицензированием; обеспечение полного контроля над данными. В то же время обращается внимание на то, что отсутствует универсальный способ генерации, применимый для любых задач, то есть в каждом конкретном случае необходимо дополнительное исследование требований, накладываемых на генерируемые данные; нет универсальных метрик качества и применимости генерируемых данных; без целенаправленного вмешательства данные получают излишне «стерильными», не позволяя учесть крайние случаи, которые возможны в реальности.

Способы увеличения объема данных на базе уже имеющихся

Сегодня МО можно считать отдельной производственной отраслью в качестве одной из форм развития систем искусственного интеллекта, целью которого является частичная или полная автоматизация решения сложных

специальных задач в самых разных областях человеческой деятельности. В то же время развитие этого направления работ идет больше в практическом, чем в научно-теоретическом плане, в связи с чем наиболее актуальные публикации, освещающие эту сферу, в том числе в формате научных статей, появляются преимущественно не в научных журналах, а в интернете на площадках обсуждения результатов практиками.

При рассмотрении любых информационных данных следует отметить, что в самом общем случае можно выделить три типа их представления, которые накладывают ограничения на применяемые способы их обработки при генерации синтетических данных. Данные представляются в формате числовых массивов, в виде текста, а также в форме изображений, в том числе видео- или аудиофайлов. В зависимости от решаемой задачи и имеющегося набора данных для МО используются данные как одного из названных типов, так и их комбинации. В то же время технологии обработки и генерации синтетических данных хотя и носят в общем характер, направленный на увеличение относительного объема целевых данных в обучающей выборке, имеют существенные различия в зависимости от формата их представления.

Так, например, для данных в формате изображений на основе одного рисунка можно сформировать некоторую их совокупность путем формальных геометрических преобразований исходного, включая такие операции, как отражение, смещение, растяжение, поворот, изменение цвета, размытие, внесение помех (штрихов, царапин, клякс), а также сочетания их в произвольных комбинациях (см. [4-8]). Характерный пример такой переработки, когда из одного натурального изображения произведено почти полсотни искусственных, представлен на рис. 1.

Понятно, что для текстовых представлений формальные преобразования противопоказаны. В этом случае необходимо заботиться прежде всего о сохранении смыслового содержания текста, и потому аугментация текстов включает другой набор операций: перестановку фраз или слов, замену их синонимами или аналогами, включение или исключение слов, применение или распаковка аббревиатур и сокращений, внесение опечаток в текстовые фрагменты, соединение и разложение текстов, генерация аналогичных текстов по заданным правилам, а также перефразировка, сохраняющая исходный смысл при заметном изменении словесного состава и характера построения фраз (см. [9-23]). Эти операции также можно применять по отдельности или в различных комбинациях, что позволяет на небольшом фрагменте текста с определенным содержанием выстроить достаточно большой объем различных текстов с аналогичным содержанием.

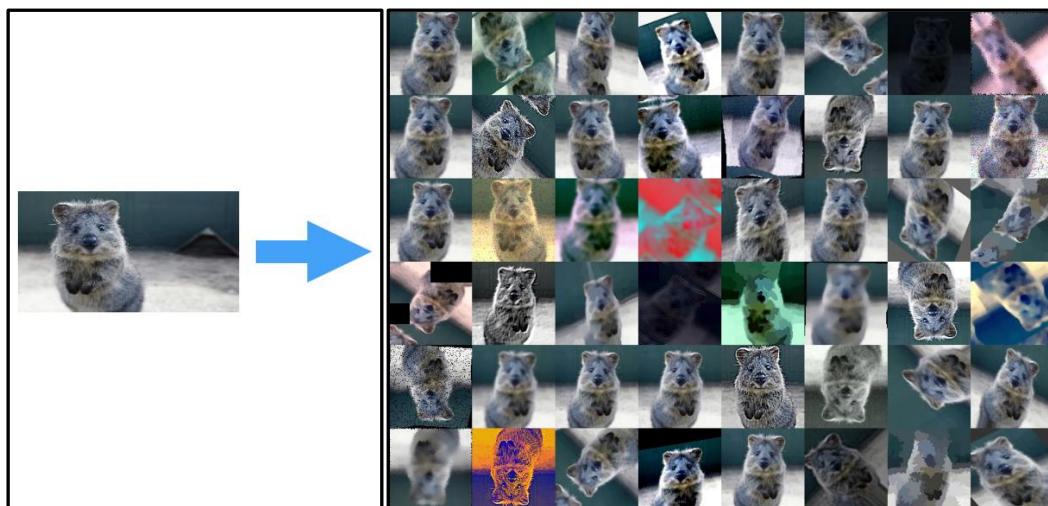


Рис. 1. Пример аугментации изображения для обучающей выборки (с сайта https://miro.medium.com/max/3564/1*bqNylp7FcqIBWg0DrcimUw.png).

При работе с числовыми массивами их элементы представляются в форме таблиц, векторов, факторов, матриц, фреймов. И, соответственно, способы обработки также имеют свою специфику, в число которых входят операции заполнения пропусков, обработки выбросов, заполнения интервалов, разномножения значений, выбраковки или дистилляции, генерации на основе моделей рассматриваемых явлений и др. (см. [24-27]).

Хотя здесь перечислено небольшое число операций, которые могут применяться к тем или иным форматам данных, следует иметь в виду, что каждая из них может выполняться различными способами. Более того, число как операций, так и способов их применения заметно шире приведенного здесь, и эксперименты в этой области продолжаются. В настоящей работе детальные особенности разнообразных способов аугментации подробно не рассматриваются, поскольку различные их вариации широко представлены в литературе (см. [4-27]).

Способы получения исходных данных по основным видам угроз на основе экспертных оценок

Выше кратко акцентировано внимание на методах создания массивов синтетических данных, использующихся специалистами по машинному обучению в приложении к различным типам данных. Все они в той или иной мере ориентированы на использование некоторой предварительной информации как о сфере машинного моделирования, так и о свойствах исходных данных, которые служат образцом для формирования обучающего множества синтетических данных. Несмотря на достаточно широкий спектр разработанных методов и технологий, специалистами-практиками подчеркивается необходимость индивидуального подбора методов под решение каждой конкретной задачи. Подчеркивается, что универсальных подходов и критериев

оценки, гарантирующих получение требуемого результата, для формирования обучающих массивов синтетических данных ни по качеству данных, ни по их количественным параметрам в настоящее время не существует. Создаваемые в процессе успешного обучения элементы ИИ ориентированы на автоматизацию решения некоторых узкоспециализированных задач. А сам процесс подбора архитектуры ИИ и способов его обучения сводится к опирающемуся на опыт и интуицию исследователя некоторому творческому процессу. При этом какие-то методы проявляют себя лучше в одной ситуации, какие-то – в другой. Специалисты отмечают, что состав обучающей базы может влиять на качество получающейся системы распознавания больше, чем любые другие факторы. Из выше сказанного можно сделать базовый вывод – в настоящее время (по крайней мере - пока) нельзя разработать универсальный «искусственный мозг», способный оперативно и с равным успехом решать произвольные задачи в сфере обеспечения безопасности населения по защите от угроз природного, природно-техногенного и антропогенного характера. Поэтому для решения множества различных задач целесообразно обучение соответствующего множества средств ИИ.

В условиях сформулированных ограничений роль экспертов должна сводиться, прежде всего, к решению следующих вопросов:

- выделению спектра проблемных задач, решение которых требует создания средств ИИ для оперативной оценки обстановки и выявления уровня рисков по каждому виду угроз;
- определению перечня анализируемой с помощью этих средств ИИ информации и источников её получения, состава и формата обрабатываемых данных;
- формулированию правил формирования массивов синтетических данных, в том числе за счет расширения границ вероятного изменения параметров в составе предоставляемой информации за пределы реально наблюдавшихся ранее.

Оценим, какого рода информация может быть использована в качестве исходной применительно к сфере оценки и моделирования основных видов угроз безопасности населенных пунктов, и каким образом ее можно получить.

Представляется, что в самом общем виде имеющиеся на сегодня возможности по получению исходной информации об угрозах безопасности населения для формирования обучающих множеств синтетических данных можно свести к следующим группам методов:

- обобщение данных натурных измерений, в том числе косвенных;
- математическое моделирование физических явлений и процессов;
- имитационное моделирование социо-биологических явлений и процессов;
- когнитивно-факторное моделирование взаимосвязей;
- использование сюжетов гипотетических представлений.

Ниже детализируется содержание этих формулировок с акцентами как на возможном составе данных, так и на их потенциальных источниках.

Интеграция данных натуральных наблюдений

Наиболее детальные и объемные исходные натурные данные для формирования обучающих массивов можно получить из специализированных мониторинговых систем.

В первую очередь, как обучающие системы, так и, впрочем, создаваемая с применением элементов ИИ система прогнозирования ЧС должны иметь доступ к данным наблюдений за метеорологической и гидрологической ситуацией в ближней и дальней окрестностях рассматриваемого населенного пункта (рис. 2). Эти данные играют

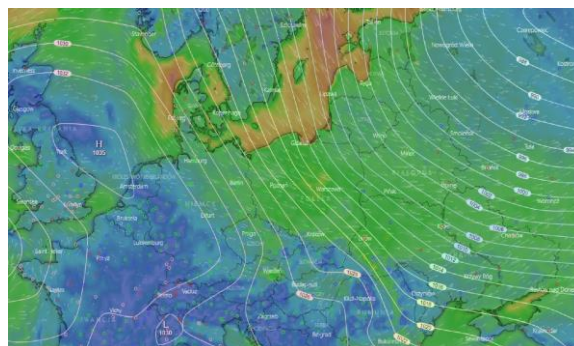


Рис. 2. Карта метеобстановки

немаловажную роль при оценке угроз наводнений, подтоплений, пожаров, распространения вредных примесей в атмосфере и водной среде, а также проявления вторичных эффектов при различных техногенных авариях. Текущие погодные условия практически всегда оказывают какое-то влияние на провоцирование, развитие или эффекты проявления различных рисков при актуализации угроз практически в любых чрезвычайных ситуациях. Например, очевидна роль интенсивных дождей в повышении уровня угрозы наводнений или в снижении уровня угрозы пожаров. Известно также, что ураганный ветер может спровоцировать аварии на сетях энергоснабжения, создающие, в свою очередь, проблемы в городских системах водо- и теплоснабжения, которые могут усугубляться низкими температурами. Однако, ровно такие же погодные условия явно не будут способствовать проведению массовых митингов и шествий.

Поэтому в массивах данных при обучении систем оценки обстановки и прогнозирования рисков целесообразно наряду с другими использовать реальные или сформированные на их основе данные метеорологической обстановки. При создании обучающих массивов для обеспечения работоспособности моделей в экстремальных погодных условиях, выходящих за рамки обычных, текущие реальные данные могут быть дополнены искусственно вводимыми значениями в рамках задаваемых экспертами диапазонов отклонений.

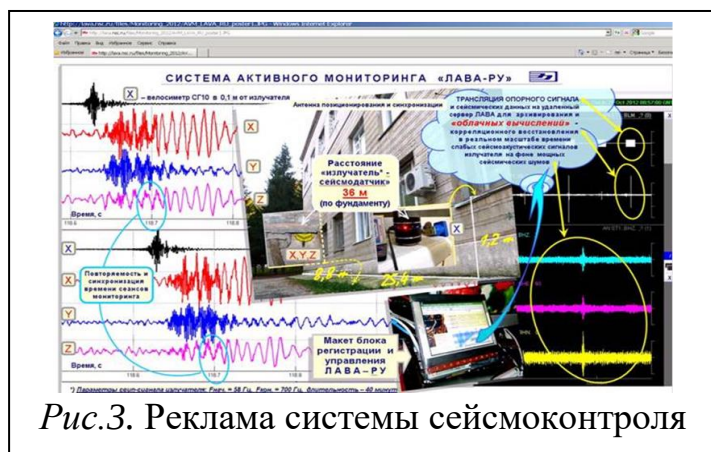


Рис.3. Реклама системы сейсмоконтроля

Подобную же роль играют данные системы сейсмического контроля (рис.3). Наиболее значимы эти данные при оценке угроз землетрясений, а также в местах, где присутствуют угрозы схода лавин, формирования селей и обвалов, прорыва плотин и т.п. Помимо предупреждения о землетрясениях в отдельных случаях именно сейсмические колебания могут стать

первичными вестниками техногенных аварий, сопровождающихся взрывами, обрушениями или прорывами. За счет фиксации и своевременной обработки таких сигналов можно повысить оперативность реагирования на техногенные ЧС. В повышении качества правильного определения параметров источника сейсмического сигнала и оценки обстановки большое значение принадлежит методам дешифровки записей сейсморегистраторов. Обычно такие сигналы за счет множественных отражений в земной коре имеют сложную структуру, «разбираться» с которой также обучают нейросети.

И, наконец, третью группу источников мониторинговых данных составляют специализированные системы контроля состояния важных технологических систем жизнеобеспечения, системы контроля окружающей среды в местах расположения экологически опасных объектов, если, конечно, они таковыми оснащены, а также информационные источники системы санитарно-эпидемиологического контроля.

В зависимости от вида угроз главенствующая роль принадлежит данным одной из этих трёх групп, данные других при этом обеспечивают сопутствующий фон обстановки.

Перечисленные выше информационные источники позволяют формировать массивы данных в числовом формате. Полученная информация может быть дополнена данными в формате текстовых документов и в формате изображений.

В качестве дополнительных исходных натуральных данных в текстовом формате целесообразно использовать различные сообщения, доклады, описания со стороны представителей дежурных служб или других свидетелей событий, накапливаемые в информационных базах или извлекаемые из других источников, находящихся в доступе (вплоть до социальных сетей).

Прямые данные в формате изображений, как правило, могут быть извлечены из архивов записей систем видеонаблюдения, которые в настоящее время присутствуют практически повсеместно в населенных пунктах и на основных трассах, а также из видеофайлов, размещаемых в социальных сетях множеством свидетелей событий.



Рис. 4. Фотофиксация урагана

Гаджеты – технические средства индивидуального использования – создали условия и породили моду среди значительной части их владельцев на создание персонального информационного контента. Распространённым явлением стала фиксация на видеокамеру смартфона каких-то событий, привлекающих внимание людей, с последующей

публикацией видео в общедоступных интернет-сервисах (рис. 4). Просматривая новости о каких-либо групповых акциях на улицах городов, нельзя не заметить, что количество любопытствующих с видеокамерами и смартфонами иногда едва ли не превышает количество участников мероприятия. В качестве еще одного примера можно вспомнить, что даже такое редкое и весьма краткосрочное явление, как падение Челябинского метеорита, было зафиксировано камерами видеонаблюдения и автомобильных видеорегистраторов, записи которых в последующем появились в свободном доступе (рис. 5).

Таким образом, за счет любителей формируется дополнительная база натуральных наблюдений, позволяющая расширить представления о происходящих явлениях или событиях, причем под различными углами зрения. Массовость видеоконтента с натурными исходными данными создает условия для формирования на их основе путем аугментации больших и разнообразных массивов синтетических данных.

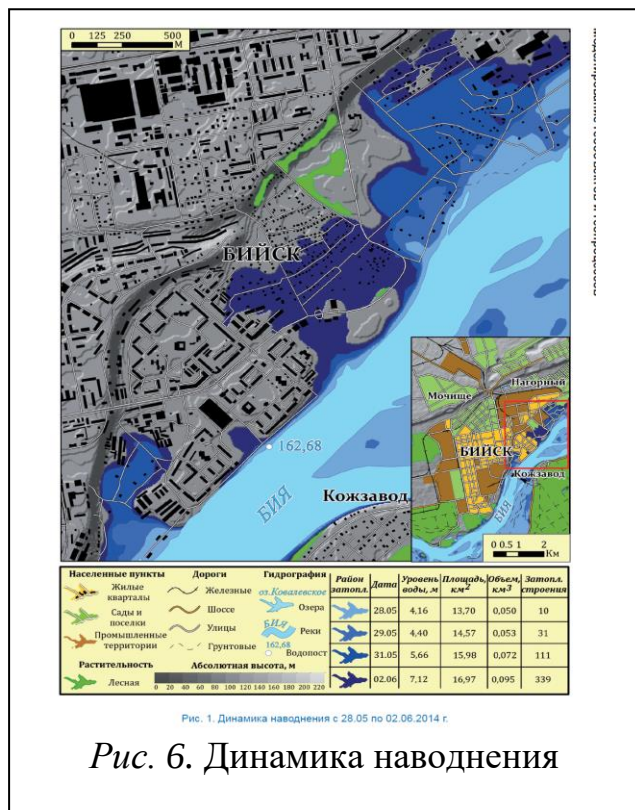
Кроме этого, получаемые в числовом формате данные часто сопровождаются их наглядным отображением в виде графиков, схем, диаграмм, которые тоже могут служить дополнительным источником данных для формирования обучающих массивов.



Рис. 5. Снимок падения метеорита

Одним из примеров использования визуализированных данных для

извлечения числовой информации является представленная в работе [28] технология использования материалов дистанционного зондирования Земли (ДЗЗ) и космических снимков. Показано, что для моделирования областей затопления в ГИС могут успешно использоваться цифровые модели рельефа (ЦМР), полученные фотограмметрическим способом по материалам дистанционного зондирования Земли (ДЗЗ). Среди всего многообразия данных ДЗЗ в настоящее время все большую роль играют космические снимки высокого и сверхвысокого разрешения. В данной работе на их основе определены, например, параметры динамики затопления прилегающих к реке зон города (рис. 6).



Расчеты по методикам инженерного уровня

Данные прямых измерений мониторинговых систем могут быть дополнены данными расчетов на базе так называемых «методик инженерного уровня». Хотя в какой-то мере такие данные и являются искусственными, но их можно считать производными от натурных, поскольку в большинстве своем подобные методики разрабатываются на основе обобщения эмпирических данных, полученных в ходе натурных наблюдений, подкрепляемых данными, полученными в ходе проведения специальных экспериментов. Под последними подразумеваются, например, результаты измерений параметров среды в целевых экспериментах по изучению особенностей горения, испарения и переноса в воздушной или водной среде различных химически-опасных веществ и их производных, которые проводились на специализированных полигонах профильных организаций. Подобные же данные касательно условий разрушения сооружений при сейсмических колебаниях были получены в ходе экспериментов на полигонах строительных или инженерных служб. При этом инженерные методики предназначены для оценки некоторой совокупности интегральных показателей, которые доступны для измерения. В качестве примера подобных методик инженерного уровня можно привести документы [29-35].

Натурные данные составляют основу для понимания сути происходящих процессов, в том числе предшествующих чрезвычайному ходу событий. Именно на их основе должны вырабатываться требования к синтетическим данным для формирования обучающих массивов. При работе с натурными данными роль экспертов сводится к определению состава требующихся для машинного обучения данных, выбору методов их аугментации для формирования обучающих массивов, включая оценку качества и характер получения дополнительных расчётных данных с помощью инженерных методик.

Математическое моделирование как источник искусственных данных

Данные натурных наблюдений служат основой для текущей оценки обстановки и своевременного получения сигналов о возрастании угроз чрезвычайных ситуаций. Однако можно ожидать, что совокупности имеющихся в наличии данных далеко не достаточно для качественного машинного обучения по следующим причинам. С одной стороны, постоянно действующие системы мониторинга включают регистрацию лишь ограниченного числа параметров, причем не обязательно те, по которым можно однозначно диагностировать возрастание риска угроз, с другой стороны, в силу крайней редкости самих ЧС полученные в этих эпизодах данные не охватывают большинства возможных вариантов развития событий.

Поэтому основным инструментом исследования динамики параметров актуализации тех или иных угроз безопасности населению в настоящее время является математическое моделирование процессов, происходящих в окружающей среде или в обществе. С помощью математической модели, описывающей закономерности взаимодействия различных сущностей, возможно изучение динамики развития нестандартных ситуаций путем задания соответствующих начальных и граничных условий. Эти подходы в последние десятилетия являются основными источниками предиктивной информации о возможных техногенных авариях и природных катастрофах, накоплен опыт применения таких моделей, ведется большая научная и практическая работа по их совершенствованию, повышению многоплановости, точности и адекватности моделей применительно к различным условиям и факторам. Однако, несмотря на постоянное параллельное наращивание возможностей и производительности вычислительных систем, как уже отмечено нами ранее, высококачественные численные модели слабо применимы в конкретных условиях ЧС как по причине высокой вычислительной сложности, так и из-за большой неопределенности в оперативном выявлении конкретных условий реализации ЧС для задания входных параметров модели.

В то же время, именно подобные математические модели могут послужить основой для формирования обучающих массивов искусственных данных путем «проигрывания» множества различных вариантов развития процессов и

событий, условия которых задаются на основе экспертных оценок вероятных происшествий. В открытом доступе находится большое число публикаций с изложением методических подходов к моделированию подобных явлений, рассмотрением разнообразных аспектов и уточняющих деталей, исследованием роли и значимости последних. В качестве примера применительно к некоторым видам угроз ниже приведены ссылки на ряд работ с описанием численных математических моделей, которые, с одной стороны, являются демонстрацией подходов исследователей к учету особенностей происходящих явлений, с другой, могут послужить инструментом формирования расширенных массивов синтетических данных применительно к различным условиям актуализации угроз.

При отборе этих примеров принималось во внимание то обстоятельство, что не всякие математические модели, предназначенные для использования в системах предупреждения и реагирования на ЧС, применимы для поставленной цели. Поскольку речь идет о машинном обучении интеллектуальных элементов систем поддержки принятия решений, которые должны быть ориентированы на возможность оперативной оценки обстановки по каким-то отрывочным сведениям и наблюдаемым на практике признакам с последующим прогнозированием вариантов развития ситуаций, то необходимо использовать модели, в которых могут воспроизводиться те самые наблюдаемые признаки и параметры проявления негативных воздействий.

Моделирование атмосферного переноса

В статье [36] приведены результаты разработки непрерывной математической модели движения многокомпонентной воздушной среды, на базе которой рассчитывается множество взаимосвязанных процессов, включая такие факторы, как транспорт загрязняющих веществ и тепла; влияние растительного покрова; изменение коэффициента турбулентного обмена; переход воды из жидкого в газообразное состояние; осаждение вещества; изменение температуры за счет конденсации и испарения аэрозоля; турбулентное перемешивание многокомпонентной воздушной среды; теплообмен между жидкими и газообразными состояниями веществ; наличие распределенных источников вещества и температуры; сила Архимеда; тангенциальное напряжение на границах раздела сред; переменная плотность, зависящая от концентрации загрязняющих веществ, температуры и давления; сжимаемость среды за счет: изменения температуры, испарения и конденсации жидкости, изменения давления, наличия источников. Констатируется разработка параллельных расчетных алгоритмов, реализованных в виде комплекса программ для проведения расчетов на высокопроизводительных вычислительных системах. Разработан и реализован комплекс программ

«AeroEcology», учитывающий процессы тепломассообмена в прибрежной зоне воздушной среды и предназначенный для построения турбулентных потоков поля скорости многокомпонентной воздушной среды на сетках с высокой разрешающей способностью, а также для расчета концентрации примесей и транспорта тепла. Пример визуализации расчета по данной модели приведен на рис.7.

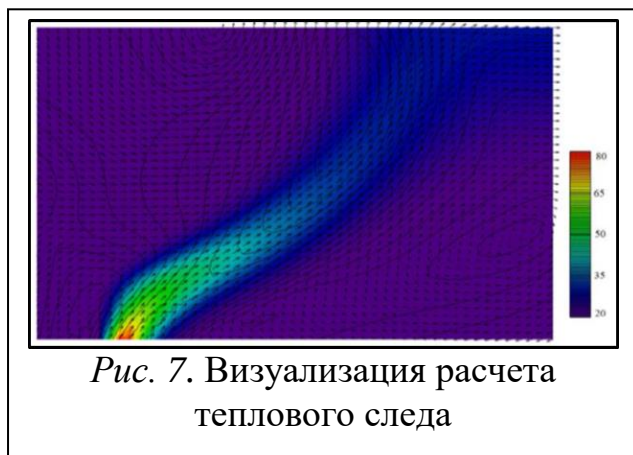


Рис. 7. Визуализация расчета теплового следа

источник загрязнения. Второй тип источника - загрязнение от локального кратковременно действующего источника (например, аварийный выброс). В этом случае через растительный массив проходит облако интенсивного загрязнения, имеющее конечные размеры. Третий тип – постоянно действующий наземный источник с локальными выбросами (например, транспортная трасса). Для такой ситуации предлагается учитывать нестационарность рассматриваемого процесса и случайный характер появления локальных загрязнителей.

Диссертация Беликова Д. А [38] посвящена разработке математической модели переноса примеси и построению эффективной параллельной реализации численного решения пространственных нестационарных адвективно-диффузионно-кинетических уравнений для исследования распространения и турбулентного рассеяния химически реагирующих загрязнителей в приземном слое атмосферы над урбанизированной территорией. Особенностью модели является математическая формализация переноса примеси с учетом химических и фотохимических реакций и модели однородного атмосферного пограничного слоя. Автором разработана модель турбулентности, включающая уравнения для дисперсии турбулентных пульсаций температуры и корреляции пульсаций температуры с пульсациями концентраций, а также алгебраические соотношения для турбулентных напряжений и потоков.

В статье Бондарева И. Н с соавторами [27] непосредственно акцентируется внимание на необходимости и возможности формирования наборов синтетических данных путем имитационного пространственного моделирования процессов распространения загрязнений для сложных условий городской застройки. С этой целью авторы апробируют применение технологии клеточных

автоматов, по их мнению, наиболее приемлемой в подобной геометрии пространства, а также приводят получаемые расчетным путем результаты.

Моделирование наводнений

Непосредственно на сайте ИВП РАН представлены материалы по применению программных комплексов для численного моделирования волн паводков и расчетов затопления территорий [39]. Отмечается, что численное гидродинамическое моделирование распространения волн паводков и наводнений различного генезиса (таяние снега, интенсивные осадки, прорыв плотин) уже выполнено для многих десятков водных объектов и гидротехнических сооружений по всей России. Разработанные ИВП РАН модели (Stream, Flood) и оригинальные расчетные алгоритмы хорошо адаптированы под реальные гидрологические ситуации. В комплекс Stream заложены возможности учета влияния ветра, осадков, переменной по времени шероховатости, фракционного состава грунтов. Нестационарные течения описываются на основе двумерных уравнений Сен-Венана, переноса примеси и тепла, транспорта неоднородных наносов и донных деформаций на основе трехслойной модели. Численное решение проводится методом типа Годунова первого и второго порядка на неструктурированных треугольно-четырёхугольных сетках в областях со сложной геометрией границы и формой подстилающей поверхности.

В основу математических моделей программного комплекса Flood положены двумерные уравнения Сен-Венана с учетом конвективных слагаемых и уравнения переноса примеси и деформаций дна в приближении мелкой воды. Для расчета паводковых русло-пойменных потоков применяется оригинальная двухслойная модель, в которой течение в руслах считается одномерным (нижний слой), а пойменное течение двумерным (верхний слой), с учетом обмена массой и импульсом между слоями. Расчет ведется на неструктурированных треугольных сетках. Границы области течения в плане и поверхность дна могут быть произвольными, возможен учет различных дорожных, берегозащитных, водопропускных и гидротехнических сооружений (дорог, мостов, дамб обвалования, полузапруд, водозаборных ковшов и т.п.).

В статье [40] изложен способ модельных расчетов волновых процессов в мелководных водоемах, альтернативный традиционному способу описания водной среды с помощью уравнений Навье-Стокса. Приводятся результаты применения программной реализации математической модели гидродинамики и сгонно-нагонных явлений в мелководных водоемах на основе регуляризированных уравнений на параллельной вычислительной системе с общей памятью на базе алгоритмов библиотеки OpenMP.

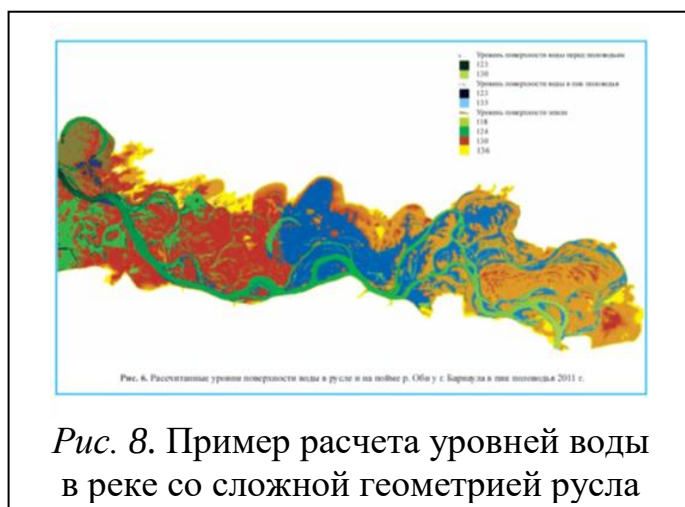


Рис. 8. Пример расчета уровней воды в реке со сложной геометрией русла

Авторы [41] приводят краткое описание нестационарной плановой модели течения для случаев сложной морфометрии русла и поймы, а также технологию численной реализации математической 2DH-модели, направленной на расчеты реальных течений (рис.8).

Для рассмотрения катастрофических ситуаций может представлять интерес работа [42], в которой обсуждаются различные

подходы по моделированию прорывов плотин и затоплению условной городской застройки с обтеканием строений волной. Рассмотрение особенностей этой задачи предлагается с помощью численного решения многоскоростного дискретного уравнения Больцмана (DBE), – как утверждается, недавно разработанного метода для моделирования транскритических потоков мелкой воды. Основным преимуществом указанного метода является простота математической модели: набор линейных, несвязанных, чисто адвективных уравнений. В этой статье DBE был оценен как метод решения уравнения мелкой воды (SWE), когда рассматриваются потоки, разрушающие плотины и воздействующие на препятствия. Такие потоки сильно нестационарны и представляют собой частые субсверхкритические и суперкритические (т.е. гидравлические скачки) переходы. Авторы подчеркивают, что исследование поведения разрушающих плотины потоков, воздействующих на препятствия, может быть полезным при моделировании процессов затопления городов. Во всяком случае, предложенная методология может с успехом применяться для генерации искусственных данных в интересах подготовки обучающих массивов с целью прогнозирования катастрофических затоплений.

Моделирование лесных пожаров

В препринте [43] предложена объемная по своему математическому описанию двумерная трехфазная математическая модель лесных пожаров, построенная методом осреднения исходных уравнений многофазной среды по высоте слоя лесных горючих материалов. Модель описывает распространение фронта лесного пожара в условиях неоднородного распределения горючих материалов на местности и наличия препятствий для распространения огня, а также учитывает перенос и выпадение горящих частиц ветром над лесным массивом. В предлагаемой модели лес рассматривается как одноярусная трехфазная среда, состоящая из воздуха и легких летучих продуктов пиролиза и горения (газовоздушная фаза), из лесных горючих материалов и твердых продуктов пиролиза лесных горючих материалов (твердая фаза) и из

относительно крупных горящих дисперсных частиц (дисперсная фаза), также являющихся продуктом пиролиза. Газовая фаза состоит из шести компонент: горючего газа CO , окислителя O_2 , углекислого газа CO_2 , паров воды H_2O , азота N_2 и мелкодисперсной сажи. При этом предполагается, что частицы мелкодисперсной сажи движутся вместе с газовой фазой. Твердая фаза также является многокомпонентной и состоит из лесных горючих материалов и продуктов их пиролиза: коксика и золы. Коксик и мелкодисперсная сажа практически полностью состоят из углерода и сгорают без твердого остатка. Дисперсная фаза в форме монодисперсной смеси полагается состоящей из горящих частиц углерода, которые переносятся ветром над лесом и оседают под действием силы тяжести. В препринте анонсирована численная реализация построенной модели на многопроцессорных вычислительных комплексах с применением метода расщепления системы уравнений по физическим процессам и конечно-разностной аппроксимацией полученных подсистем.



Рис. 9. Динамика пожарной обстановки

Перминов В.А. [44] использует систему уравнений в сходной постановке задачи. В своей работе приводит результаты модельных исследований особенностей процессов возгорания, распространения горения в форме низовых или верховых пожаров, особо выделяет условия перехода низовых пожаров в верховые. (рис. 9)

В диссертационной работе Барановского Н.В. [45] основной упор сделан на изучение и модельное воспроизведение процессов возгорания лесных массивов от различных источников тепловой энергии (разрядов молний, фокусирования солнечного излучения, горящих частиц).

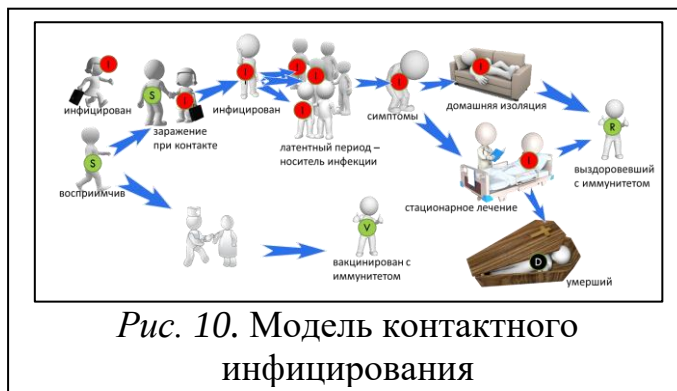
Математическое моделирование низовых лесных и степных пожаров рассматривается в диссертации Бурасова Д.М. [46], в которой представлены упрощенная постановка задачи с выделением основных влияющих факторов и выведены аналитические зависимости для оперативных оценок.

Моделирование процессов в социальных системах

Если рассмотренные выше модели процессов, протекающие в физических средах, базируются на фундаментальных законах сохранения массы, энергии и импульса, то в моделях описания поведения индивидуумов или их групп в социальных системах применяются представления, базирующиеся на статистических наблюдениях, а в ряде случаев просто на некоторых экспертных представлениях. В настоящее время методы математического моделирования социальных систем находятся в стадии интенсивного развития.

Одним из перспективных с точки зрения применения в целях формирования синтетических данных для обучающих массивов, на наш взгляд, является метод агентного моделирования.

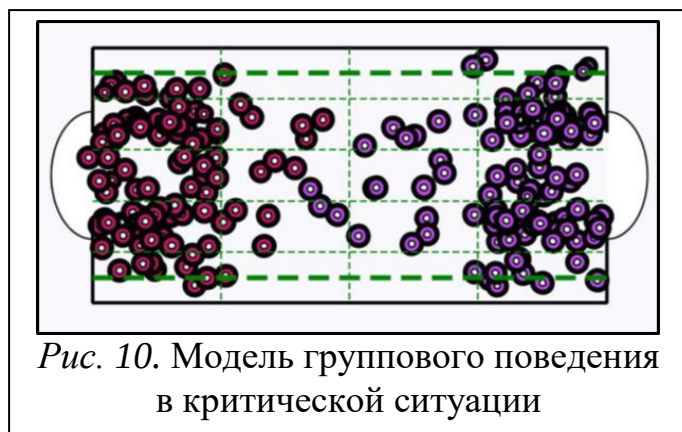
Например, М. Кондратьев в своей диссертационной работе [47] предложил дискретно-событийную многоагентную стохастическую модель распространения гриппа А на территории города, параметры которой могут быть настроены для моделирования не только гриппа, но и других видов инфекционных заболеваний. Модель позволяет воспроизводить цепочки



контактных взаимодействий, при которых с какой-то вероятностью у агента может изменяться состояние здоровья и иммунитета, то есть переход из состояния «здоровый» (S) в состояние «вакцинированный» (V) или «инфицированный» (I) с последующим выздоровлением (R) или смертью (D) (рис.10).

Подобные принципы мультиагентного моделирования инфекционной эпидемии легли в основу разработки полномасштабной модели, создаваемой применительно к отечественным реалиям в рамках Соглашения №075-11-2020-011 от 19.10.2020 между консорциумом исполнителей и Минобрнауки по теме «Моделирование эпидемий вирусных инфекций» при головной роли РФЯЦ ВНИИТФ. После завершения разработки эта модель, на наш взгляд, будет наиболее адекватной для ее применения в заявленных целях, в том числе в целях получения разнообразных синтетических данных для гипотетических сценариев инфекционных эпидемий.

Аналогичные подходы развиваются в настоящее время для изучения поведения групп населения в других кризисных ситуациях.



Так, Бекларян А.Л. в своей работе [48] разработал агентную модель поведения толпы в условиях чрезвычайной ситуации с учетом таких эффектов, как давка, паника, эффект турбулентности толпы и возникновение волн сжатия. На ее основе можно исследовать динамику группового поведения людей при различных значениях внешних управляющих параметров.

Для повышения временной эффективности системы, в данную модель имплементирован предложенный эволюционный (модифицированный)

алгоритм нечеткой кластеризации, позволивший перейти от динамики агентов к динамике кластеров и изучить процесс выявления устойчивых групп соответствующих агентов.

В статье [49] приводится изложение мультиагентной модели, предназначенной для изучения процесса эвакуации людей из помещений при чрезвычайных ситуациях, которая учитывает сложную конфигурацию ограниченного пространства, влияние стен и других препятствий, физическое взаимодействие людей в процессе эвакуации и разумную взаимную корректировку их целей.

Свой вариант решения аналогичной задачи также на базе агентного подхода предложил Гребенников Р. А. в диссертационной работе [50]. Им разработан оригинальный метод применения скалярных полей в зоне нахождения агента для определения параметров его свободы перемещений в заданном направлении. В ходе выполнения темы разработаны также алгоритм и программа расчета поведения толпы как многочисленной группы агентов при различных условиях обстановки.

Все перечисленные программные комплексы и их многочисленные аналоги могут быть использованы для генерации синтетических данных с целью создания обучающих множеств в соответствии с задачами, сформулированными экспертами.

Использование косвенных данных из других информационных источников

В научной литературе приводится [51] описание результатов практического эксперимента, впервые проведенного компанией Google в 2008 году. В этом эксперименте предпринята успешная попытка по установлению возможности идентификации начала эпидемии гриппа по косвенным информационным данным, в качестве которых использовалось изменение частоты запросов определенного вида в поисковике. Сопоставление с реальными данными по развитию эпидемии позволило выделить спектр наиболее релевантных запросов, коррелирующих с интенсивностью роста заболеваемости, и обучить нейронную сеть, настроив ее на выделение и анализ соответствующей информации. Созданный программный комплекс с большой точностью позволил прогнозировать темпы и направления распространения эпидемии.

Этот подход может быть применен для выявления и косвенной диагностики множества различных изменений в обществе – от коллективного здоровья до общественных настроений или чрезвычайных происшествий. Учитывая, что в



Рис. 10. Условное представление виртуального робота

последние годы большое развитие получили технологии создания ботов – под которыми понимаются виртуальные роботы, функционирующие на основе специальных программ, с их применением возможна генерация разного рода текстовых сообщений для формирования обучающих множеств. Задачей ботов является автоматическое и/или по заданному расписанию выполнение каких-либо действий через интерфейсы, предназначенные для людей. В варианте чат-ботов они способны имитировать реальное сетевое

общение с пользователем с помощью текстовых или аудио сообщений на сайтах, в мессенджерах, мобильных приложениях или по телефону. Некоторые чат-боты оснащаются искусственным интеллектом, настраиваются на определенные темы общения и постоянно дообучаются в процессе функционирования. Подобные программы-роботы могут быть применены для генерации множества текстовых сообщений по сформированным экспертами образцам, шаблонам и правилам. Полученные путем генерации искусственные текстовые сообщения могут быть размножены методами аугментации, чтобы их можно было использовать для создания обучающих множеств синтетических сообщений. Обученные на них программы СППР будут ориентированы на своевременное выявление каких-либо косвенных признаков нарастающих угроз в информационных потоках и выработку последующего прогноза развития обстановки.

Использование искусственных видеоматериалов



Рис. 10. Кадр из фильма о стихийном бедствии

Вероятно, в какой-то мере может показаться экзотичным, однако можно считать вполне работоспособным способом дополнение редких натуральных материалов документальных видео- и фотосъемок их художественным представлением. В современной киноиндустрии много внимания уделяется изображению различных катастрофических ситуаций, в которые попадают киногерои, – от обычных автокатастроф до антропогенных аварий или стихийных бедствий. Несмотря на то,

что подобные ситуации являются в большинстве своём «придуманными», большое значение со стороны производителей этого контента придается

реалистичности их изображения. Благодаря такому отношению к выполнению своей работы со стороны деятелей киноиндустрии можно найти достаточно большое количество высококачественных искусственных изображений под множество ситуаций, которые требуются для решения задач машинного обучения СППР. Покадровое разделение видеоматериалов с последующей аугментацией полученных изображений может дать базу для формирования обучающих множеств с созданием ИИ по анализу видеопотоков.



Рис. 10. Кадр из ролика об авиационной аварии

Представляется, что аналогичные видеоданные, хотя и более условные, что, впрочем, может оказаться даже более выгодно для решения поставленной задачи, могут быть получены путем нарезки кадров из материалов игрового имитационного моделирования различных нестандартных ситуаций и происшествий. В частности, к таковым можно отнести как собственно видеоигры с качественным представлением игрового мира, так и комментирующие какие-то

события видеоролики, в которых с помощью имитационных моделей проигрывается динамическое развитие ситуации, позволяющее провести визуальный анализ и наглядную демонстрацию особенностей происшествия. Очень часто такие видеоролики монтируются для визуального воспроизведения аварий, например, на воздушном, водном, железнодорожном или автомобильном транспорте. Как правило, в основе таких имитаций помимо имеющейся скудной информации о самом происшествии лежат инженерные расчеты, а также интерпретация того и другого со стороны опытных экспертов, что уже само по себе может служить основанием для гарантий некоторой степени достоверности получаемых на их основе данных и изображений.

Заключение

В настоящем препринте обсуждаются проблемные вопросы создания основанных на средствах искусственного интеллекта систем поддержки принятия решений в кризисных и чрезвычайных ситуациях. Основной проблемой при решении этой задачи является формирование обучающих множеств, так как в силу редкости и уникальности подобных явлений данных натурных наблюдений о подобных событиях критически мало. В препринте предложен подход, основанный на внедрении «гибридного интеллекта», органично сочетающего экспертные знания и опыт, аналитические расчетные модели, средства машинного обучения. Проведен обзор и анализ большого количества литературных источников, предложены способы и возможные источники как для получения дополнительной информации на основе натурных наблюдений, так и для формирования наборов искусственных данных, которые

могут быть использованы в качестве основы формирования обучающих множеств.

Библиографический список

1. Синтетические данные. https://en.wikipedia.org/wiki/Synthetic_data
2. СИНТО. Платформа синтетических данных. Что такое синтетические данные? <https://www.syntho.ai/ru/what-is-synthetic-data/>
3. V7. What is Synthetic Data in Machine Learning and How to Generate It. <https://www.v7labs.com/blog/synthetic-data-guide>
4. SmartEngines. Аугментация (augmentation, “раздутие”) данных для обучения нейронной сети на примере печатных символов <https://habr.com/ru/company/smartengines/blog/264677/?fl=ru,en>
5. Zhang H. Mixup: Beyond Empirical Risk Minimization / Zhang H., Ciss M., Dauphin Y., Lopez D. // ArXiv e-prints [Электронный ресурс] – ArXiv:1710.09412 – 2017 – URL: <https://arxiv.org/abs/1710.09412>
6. Chernyshova Y. Generation method of synthetic training data for mobile OCR system / Chernyshova Y. Gayer A. Sheshkus A. // Proc. SPIE 10696, Tenth International Conference on Machine Vision (ICMV 2017), 106962G (13 April 2018); doi: 10.1117/12.2310119
7. Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, Alexei A. Efros. Dataset Distillation. <https://arxiv.org/abs/1811.10959>
8. Geoffrey Hinton, Oriol Vinyals, Jeff Dean. Distilling the Knowledge in a Neural Network. <https://arxiv.org/abs/1503.02531>
9. Xiang Zhang, Junbo Zhao, Yann LeCun. Character-level Convolutional Networks for Text Classification. arXiv:1509.01626v3 [cs.LG] 4 Apr 2016. <https://arxiv.org/pdf/1509.01626.pdf>
10. Jurafsky, Daniel; H. James, Martin (2000). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J.: Prentice Hall. ISBN 978-0-13-095069-7. <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>.
11. William Yang Wang and Diyi Yang. 2015. That’s So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2557–2563, Lisbon, Portugal. <https://aclanthology.org/D15-1306/>
12. Marzieh Fadaee, Arianna Bisazza, Christof Monz. Data Augmentation for Low-Resource Neural Machine Translation. <https://arxiv.org/pdf/1705.00440.pdf>
13. Sosuke Kobayashi. Augmentation: Data Augmentation by Words with Paradigmatic Relation. <https://arxiv.org/pdf/1805.06201.pdf>
14. Siddhant Garg, Goutham Ramakrishnan. BAE: BERT-based Adversarial Examples for Text Classification. <https://arxiv.org/pdf/2004.01970.pdf>

15. Bidirectional Encoder Representations from Transformers BERT (language model). [https://en.wikipedia.org/wiki/BERT_\(language_model\)](https://en.wikipedia.org/wiki/BERT_(language_model))
16. Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, Quoc V. Le. Unsupervised Data Augmentation. <https://arxiv.org/pdf/1904.12848.pdf>
17. Hanjie Chen, Yangfeng Ji. Improving the Explainability of Neural Sentiment Classifiers via Data Augmentation. <https://arxiv.org/pdf/1909.04225.pdf>
18. Marzieh Fadaee, Arianna Bisazza, Christof Monz. Data Augmentation for Low-Resource Neural Machine Translation. <https://arxiv.org/pdf/1705.00440.pdf>
19. Franco M. Luque. Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis. <https://arxiv.org/abs/1909.11241>
20. Hongyu Guo, Yongyi Mao, Richong Zhang. Augmenting Data with Mixup for Sentence Classification: An Empirical Study. <https://arxiv.org/pdf/1905.08941.pdf>
21. Kushal Kafle, Mohammed Yousef Hussien, and Christopher Kanan. 2017. Data Augmentation for Visual Question Answering. In Proceedings of the 10th International Conference on Natural Language Generation, pages 198–202, Santiago de Compostela, Spain. Association for Computational Linguistics. DOI: 10.18653/v1/W17-3529
22. V. Kumar, A. Choudhary, E. Cho. Data augmentation using pre-trained transformer models. arXiv preprint arXiv:2003.02245, 2020 - arxiv.org
23. Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring Transformer Text Generation for Medical Dataset Augmentation. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4699–4708, Marseille, France. European Language Resources Association. URL: <https://aclanthology.org/2020.lrec-1.578>
24. Brett Lantz. Machine Learning with R - Third Edition/ <https://www.packtpub.com/product/machine-learning-with-r-third-edition/9781788295864>
25. Подготовка данных для алгоритмов машинного обучения. <http://blog.datalytica.ru/2018/04/blog-post.html>
26. Медведев Д., Дьяконов А. Новые свойства метода обработки данных при работе с табличными данными. <https://arxiv.org/abs/2010.09839>
27. Бондарев И.Н., Матерухин А.В., Гвоздев О.Г. Использование клеточных автоматов для имитационного моделирования распространения загрязнения атмосферного воздуха в условиях мегаполиса / Проблемы оптимизации сложных систем – 2020, с.10-15, DOI: 10.24411/9999-018A-2020-10002
28. Новаковский Б.А., Колесникова О.Н., Прасолова А.И., Пермяков Р.В. Геоинформационное моделирование наводнений по материалам космической съемки (на примере г. Бийск, Алтайский край). Геоинформатика–2015, №1, с.15-20

29. Приказ Минприроды от 6 июня 2017 г. N 273 «Об утверждении методов расчетов рассеивания выбросов вредных (загрязняющих) веществ в атмосферном воздухе», 96 с.
30. Руководство по безопасности «Методика моделирования распространения аварийных выбросов опасных веществ». Серия 27. Выпуск 11. — М.: ЗАО «Научно-технический центр исследований проблем промышленной безопасности», 2015. — 130 с.
31. Руководство по организации контроля состояния природной среды в районе расположения АЭС/ Под ред. К.П. Махонько. Л.: Гидрометеиздат. 1990. 264 с.
32. Руководство по организации контроля состояния природной среды в районе расположения АЭС/ Под ред. К.П. Махонько. Л.: Гидрометеиздат. 1990. 264 с.
33. Горский В.Г., Моткин Г.А., Петрунин В.А., Терещенко Г.Ф., Шаталов А.А., Швецова-Шиловская Т.Н. Научно-методические аспекты анализа аварийного риска. М. Экономика и информатика. 2002. 260 с.
34. Методика прогнозирования масштабов заражения СДЯВ при авариях (разрушениях) на химически опасных объектах и транспорте. РД 52.04.253–90. – Л.: Гидрометеиздат, 1991. – 24 с.
35. Количественная оценка риска химических аварий. / Под. ред. В.М. Колодкина, Ижевск, 2001, 228 с.
36. Сухинов А.И., Хачунц Д.С., Чистяков А.Е. Математическая модель распространения примеси в приземном слое атмосферы и ее программная реализация на многопроцессорной вычислительной системе / Вестник УГАТУ, 2015, т.19, №1(67), с.213-223
37. Бояршинов М. Г. Математическое моделирование переноса и рассеяния газообразных примесей в растительных массивах. Диссертация, доктор технических наук, 2018 г. 335 с.
38. Беликов Д. А. Параллельная реализация математической модели атмосферной диффузии для исследования распределения первичных и вторичных загрязнителей воздуха над урбанизированной территорией. Диссертация, кандидат физико-математических наук, 2006 г., 177 с.
39. Моделирование волн паводков и расчеты затопления территорий. <https://models.iwp.ru/modelirovanie-razvitiya-prirodnih-i-antropogennyh-navodnenij-ruslovyh-deformacij#prognozyzatopleniya>
40. Шишениа А.В., Кузнецова И.Ю., Сухинов А.И. Разработка математической модели затопления прибрежных районов на основе гиперболизированных уравнений гидродинамики // Международный журнал прикладных и фундаментальных исследований. – 2018. – № 5-2. – С. 306-311; URL: <https://applied-research.ru/ru/article/view?id=12260>
41. Зиновьев А.Т., Кошелев К.Б. Моделирование процесса затопления пойменных территорий для участков крупных рек со сложной

- морфометрией русла и поймы. Водное хозяйство России № 6, 2013. с 17-31.
42. M. la Rocca, S. Miliani, P. Prestininzi. Discrete Boltzmann Numerical Simulation of Simplified Urban Flooding Configurations Caused by Dam Break. *Front. Earth Sci.*, 27 October 2020. <https://doi.org/10.3389/feart.2020.00346>
 43. Кулешов А.А., Мышецкая Е.Е., Якуш С.Е. Двумерная трехфазная математическая модель лесных пожаров // Препринты ИПМ им. М.В.Келдыша. 2017. № 92. 12 с. doi:10.20948/prepr-2017-92 URL: <https://library.keldysh.ru/preprint.asp?id=2017-92>
 44. Перминов В. А. Математическое моделирование возникновения верховых и массовых лесных пожаров. Диссертация, доктор физико-математических наук, 2010, 282 с.
 45. Барановский Н. В. Теплофизические аспекты прогностического моделирования лесной пожарной опасности. Диссертация, доктор физико-математических наук, 2012, 436 с.
 46. Бурасов Д.М. Математическое моделирование низовых лесных и степных пожаров и их экологических последствий. Диссертация кандидата физико-математических наук, 2006. 192 с.
 47. Кондратьев М. А. Разработка модели распространения инфекционных заболеваний на основе агентного подхода. Диссертация, кандидат технических наук. 2012, 201 с.
 48. Бекларян А.Л. Агентная модель поведения толпы в условиях чрезвычайной ситуации для оценки интенсивности фронта выходного потока, диссертация, кандидат технических наук, 2016, 201 с.
 49. Самарцев А. А., Иващенко В. А., Резчиков А. Ф., Кушников В. А., Филимонюк Л. Ю., Богомолов А. С. Мультиагентная модель процесса эвакуации людей из помещений при возникновении чрезвычайных ситуаций / Управление большими системами, Вып. 72, М.: ИПУ РАН, 2018, с. 217-244
 50. Гребенников Р. А. Моделирование поведения толпы с использованием локальных скалярных полей, диссертация, кандидат технических наук, 2011, 113 с.
 51. Eisenstein M. Infection forecasts powered by big data. *Nature*. 2018 Mar 8;555(7695):S2-S4. doi: 10.1038/d41586-018-02473-5. PMID: 29517020

Оглавление

Введение	3
Понятие синтетических данных.....	6
Способы увеличения объема данных на базе уже имеющихся.....	6
Способы получения исходных данных по основным видам угроз на основе экспертных оценок	8
Интеграция данных натурных наблюдений	10
Расчеты по методикам инженерного уровня.....	13
Математическое моделирование как источник искусственных данных	14
Моделирование атмосферного переноса	15
Моделирование наводнений	17
Моделирование лесных пожаров.....	18
Моделирование процессов в социальных системах	19
Использование косвенных данных из других информационных источников....	21
Использование искусственных видеоматериалов.....	22
Заключение.....	23
Библиографический список.....	24