Оригинальная статья
https://doi.org/10.21683/1729-2646-2020-20-4-25-34

Оценка безопасности искусственного интеллекта

Йенс Брабанд[1], Хендрик Шебе[2]*
[1]Siemens Mobility GmbH, Брауншвейг, Германия, [2]TÜV Rheinland, Кельн, Германия
**schaebe@de.tuv.com*


Йенс Брабанд


Хендрик Шебе

Резюме. Цель. *В данной статье обсуждается подход к оценке безопасности систем с искусственным ителлектом (ИИ). Это актуально для тех случаев, когда ИИ используется в системах, связанных с обеспечением безопасности, а также применительно к железнодорожным системам автоматизации, в составе которых предполагается применение средств ИИ.* **Методы.** *Основное внимание в работе уделено не столько самому ИИ, сколько оценке его безопасности. Более пристальное внимание к моделям ИИ показывает, что многие из них, в особенности машинное обучение, являются статистическими. Таким образом, при проведении оценки безопасности, помимо выполнения обычных процедур, необходимо подвергнуть анализу модель, используемую в ИИ.* **Результаты.** *Часть допустимой интенсивности опасных случайных отказов, предусмотренных для соответствующего уровня полноты безопасности, должна отводиться для вероятностного сбойного поведения системы ИИ. Авторы излагают свои идеи на простых примерах и предлагают тему для научных исследований, разработка которой может сыграть решающую роль при внедрении ИИ в ответственные системы.* **Заключение.** *Представлен метод экспертизы безопасности систем с искуственным интеллектом.*

Ключевые слова: *искуственный интеллект, оценка безопасности, функциональная безопасность.*

Для цитирования: *Брабанд Й, Шебе Х. Оценка безопасности искусственного интеллекта // Надежность. 2020. №4. С. 25-34. https://doi.org/10.21683/1729-2646-2020-20-4-25-34*

Поступила *20.06.2020 г.* / **После доработки** *05.09.2020 г.* / **К печати** *18.12.2020 г.*

1. Введение

В последние годы значительно вырос интерес к искусственному интеллекту (ИИ); сфера его применения также расширилась. К ней относятся, например:

- обработка данных;
- системы поддержки принятия решений;
- распознавание речи;
- распознавание лиц;
- роботы-медсестры;
- системы автономного вождения;
- искусство и т.д.

Некоторые из сфер применения ИИ могут иметь отношение к безопасности. В связи с этим при внедрении ИИ необходимо руководствоваться стандартами функциональной безопасности [8, 9, 10], а также проводить оценку безопасности.

В настоящей статье рассматривается подход к оценке безопасности систем с ИИ. Во втором разделе дается определение термина ИИ. В третьем разделе описывается процесс определения уровня полноты обеспечения безопасности для систем ИИ. В четвертом разделе более детально рассмотрены системы ИИ в целях обеспечения лучшего понимания систем ИИ и определения подхода к обеспечению функциональной безопасности. В пятом разделе описываются требования к обеспечению функциональной безопасности систем ИИ и возможная процедура оценки. В шестом разделе приведен пример проведения оценки безопасности очень простой системы. В последнем разделе подводятся итоги проделанной работы.

2. Что такое искусственный интеллект?

Существует множество публикаций с упоминанием систем, использующих ИИ. Например, Брюнет [3] сделал краткий обзор таких систем. Отправной точкой в развитии возможностей ИИ в 1950-х гг. стал тест Тьюринга, основной целью которого была проверка, может ли машина имитировать интеллектуальное поведение, свойственное человеку. Впоследствии была разработана концепция эволюционных программ. Термин «искусственный интеллект» был впервые использован в Дартмутском колледже в 1956 году. Многими учеными были предложены и другие концепции.

ИИ можно определить как интеллект, демонстрируемый машиной. ИИ имитирует когнитивные функции, обучение, решение проблем и т.д.

Вопрос заключается в следующем: могут ли представленные ниже критерии служить критериями искусственного интеллекта:

- использование речи;
- разум;
- самоанализ.

Несмотря на то, что достижения в области развития ИИ действительно поражают, в статьях и презентациях много «хайпа» вокруг глубокого машинного обучения (см., например, статью Хэтташ и Гайслер, 2019 [7]), а вот публикаций с полноценным обоснованием безопасности применения ИИ, насколько нам известно, до сих пор нет, хотя есть немало исследовательских проектов, направленных на обоснование безопасности ИИ.

В последнее время разработан ряд подходов к обеспечению безопасности, в частности, проект стандарта UL 4600 [15], содержащий требование проведения доказательства безопасности при оценке транспортных средств, в составе технических средств которых могут использоваться алгоритмы ИИ. При этом в проекте стандарта UL 4600 подробно описан обсуждаемый предмет, но не процесс подтверждения соответствия требованиям безопасности. Это четко указано во введении: «Соответствие данному стандарту не является гарантией безопасности автоматизированного транспортного средства». Акцент делается на «воспроизводимой оценке оценки полноты доказательства безопасности». Проект стандарта UL 4600 предполагается использовать в качестве дополнения к стандарту МЭК 61508 [10].

Другие комитеты по стандартизации, например, немецкий DKE, сосредоточили внимание на процессном подходе и концепции жизненного цикла. Путцер [13] продвигает λ_{AI}, критерий, аналогичный интенсивности опасных отказов в функциональной безопасности, но не дает ему четкого определения.

3. Необходим ли SIL применительно к ИИ?

В этом разделе обсуждается вопрос о том, требуется ли определение уровня полноты безопасности применительно к ИИ, и если да, то какой должна быть соответствующая процедура.

Концепция уровня полноты безопасности (SIL) используется во многих стандартах в области функциональной безопасности. Первый среди них – широко известный стандарт МЭК 61508. В статье Шебе [14] дан метод определения SIL.

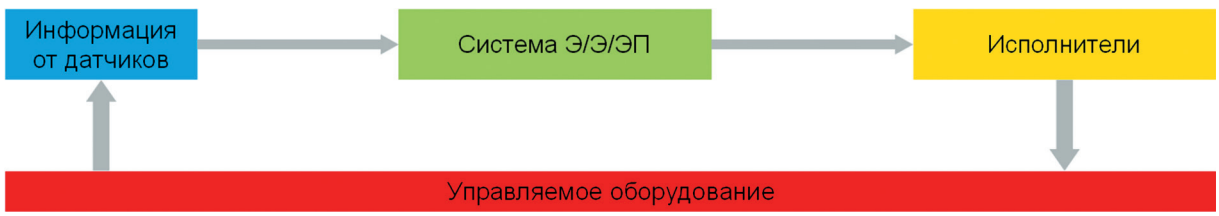

Рис. 1. Э/Э/ПЭ система управления

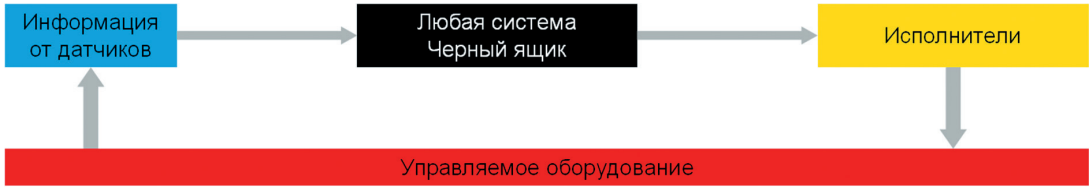

Рис. 2. Произвольная система управления (черный ящик)

На рис. 1 показана стандартная электрическая, электронная, программируемая электронная система (система Э/Э/ПЭ). На схеме изображены контролируемое оборудование, информация, поступающая от датчиков в систему управления, а также исполнительные устройства под контролем системы управления. В зависимости от последствий отклонений в работе системы управления, она получает соответствующий уровень полноты обеспечения безопасности (SIL).

В общем, тип системы управления не имеет значения. Применительно к задачам анализа степени риска и определения SIL она в любом случае рассматривается как черный ящик. Это представлено на рис. 2.

В данном случае черный ящик также может быть системой ИИ. Таким образом, может быть необходимо определение уровня полноты безопасности, если система ИИ выполняет задачи, связанные с безопасностью, а SIL может быть определен теми же методами, что и для системы Э/Э/ПЭ. Правила же определения SIL могут различаться в зависимости от типа системы, которая реализует функции черного ящика.

Какого SIL следует ожидать применительно к различным способам использования ИИ? В основном это будет зависеть от последствий отказов и от возможностей использования других мер снижения риска отказов:

- обработка данных (SIL зависит от результатов и их использования);
- системы поддержки принятия решений (обычно SIL не присваивается, если человек имеет возможность отменить решение системы);
- распознавание речи (SIL зависит от того, как используются результаты, и наличия безопасного резервирования);
- распознавание лиц (SIL зависит от того, как используются результаты, т.е. какие функции активированы);
- роботы-медсестры (поскольку такие роботы дают пациентам лекарства, помогают им передвигаться, требуется определение SIL);
- автономные системы транспорта (наличие SIL необходимо ввиду возможных дорожно-транспортных происшествий).

В любом случае должен проводиться анализ опасностей и риска, чтобы определить SIL, или обоснование отсутствия необходимости его определения. Необходимо обеспечить соблюдение положений соответствующего стандарта функциональной безопасности.

4. Искусственный интеллект изнутри

Архитектура ИИ

На рис. 3 изображена упрощенная архитектура системы ИИ. Она в чем-то схожа с архитектурой, предложенной Вангом [16], но не повторяет ее.

Система ИИ основана на использовании модели, которой свойственна гибкость и необходимость в обучении, которое проводится на основе имеющихся данных. Эти данные должны быть репрезентативны, т.е. достоверно отражать будущие события. Важно не допускать ситуаций, подобных изложенной в статье Корни [5], когда система ИИ продемонстрировала расовую дискриминацию по той причине, что в систему был внесен нерепрезентативный набор данных для обучения.

После обучения устанавливаются параметры модели. Позже она используется для создания запросов на получение данных и активации объектов для управления кон-

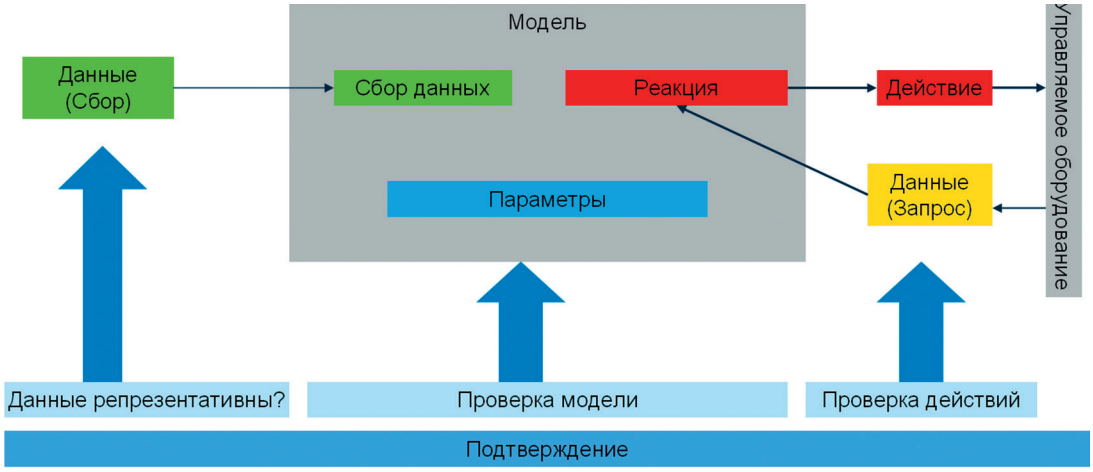

Рис. 3. Архитектура системы ИИ

тролируемым оборудованием. Продолжение обучения возможно даже после ввода системы в эксплуатацию.

Для продолжения обучения важно:

- проверить модель;
- убедиться в репрезентативности данных;
- проверить верификацию цепочки «данные – реакция модели – действия»;
- провести общую валидацию.

При верификации модели используются опытные данные, которые также должны быть репрезентативны и не могут совпадать с данными, используемыми для обучения.

В следующих подразделах мы более подробно рассмотрим несколько типов систем ИИ. Это позволит оптимизировать модельную часть архитектуры, показанной на рис. 3.

Взгляд на ИИ с использованием анализа сходства

Как видно из рис. 3, большинство алгоритмов ИИ основаны на статистике или, по крайней мере, имеют сходство с ней. В качестве первого подхода к определению требований для использования ИИ в ответственных системах может быть использована статистическая процедура. Такой метод можно также назвать анализом сходства. Какие результаты может принести статистическая процедура? Что, если алгоритмы ИИ, например, машинное обучение, интерпретировать как статистическое выравнивание данных, но с очень сложными алгоритмами и большими выборками? Необходимо отметить, что речь идет об упрощенном представлении, которое используется для общего понимания ИИ, что позволит применять методы оценки безопасности.

Чтобы понимать данный пример на интуитивном уровне, рассмотрим одну из самых простых статистических моделей, которая известна каждому инженеру еще со школьных времен – линейная регрессия, т.е. подгонка (прямой) линии к данным. Что мы можем извлечь из этого? Следует обратить внимание на то, что этот результат не новый. Перл и Маккензи уже заявили о том, что нейронные сети «…контролируются посредством проведения исследований и получения результатов, к которым инженеры пытаются подогнать функцию, почти так же, как специалист по статистике пытается подогнать прямую линию к набору точек». Но, насколько известно авторам, это сходство еще исследуется.

Допустим, что некоторое критически важное для безопасности решение будет зависеть от качества кривой, построенной по точкам. Анскомб [1] в своей статье продемонстрировал, что может случиться при построении кривой по точкам. В наборе точек Анскомба все соответствующие статистические показатели имеют, как минимум, два знака после запятой, хотя, очевидно, что наборы отличаются друг от друга (рис. 4).

На рис. 4 приведены примеры верной подгонки данных (набор данных 1); набор данных (2), где, очевидно, используется неправильная модель; набор данных (3), на который влияют выбросы; набор данных (4) с точкой выплеска, которая является результатом совершенно неверной постановки эксперимента. Даже из этого простого примера можно сделать несколько важных выводов:

1. Модель должна быть правильной, иначе мы никогда не сможем должным образом подогнать данные (см. набор данных 2), независимо от того, как долго будет продолжаться обучение или насколько качественными являются данные.

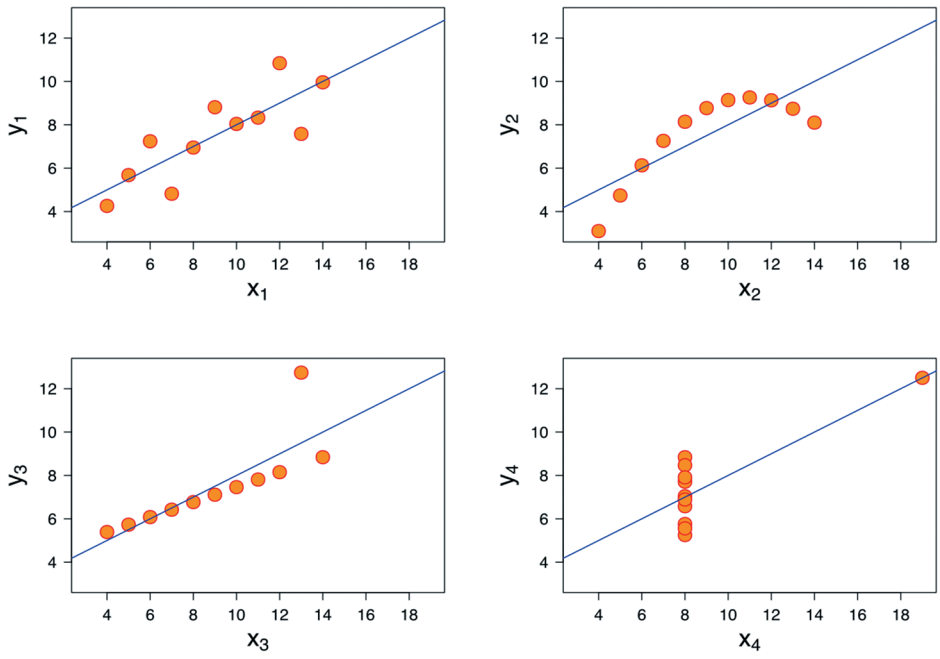

Рис. 4. Примеры – что можно извлечь из линейной регрессии

2. Набор данных для обучения должен быть репрезентативным для реальных данных; необходимо убедиться, что выборка является адекватной (см. набор данных 4).

3. Необходимо наличие метода для обнаружения выбросов (или даже для их устранения, см. набор данных 3) или аномалий.

4. Необходимо наличие критерия для определения качества подгонки (как коэффициент детерминации R2 в нормальной регрессии). Но такой критерий и рассчитанная подгонка данных зависят от функции потерь (см. набор данных 1, где обычная функция потерь с методом наименьших квадратов представлена, как и все другие подгонки, на рисунке 4).

Машинное обучение как задача классификации

Машинное обучение (МО) является наиболее успешным процессом реализации ИИ. С точки зрения статистики, МО можно интерпретировать как задачу классификации, которая обеспечивает другой взгляд на проблему. Все выводы из предыдущего раздела непосредственного относятся к МО. Большинство алгоритмов МО решают задачи классификации, аналогичные кластерному анализу или дискриминантному анализу в статистике. Имеется (как минимум) два класса (больших) данных в многомерном пространстве (см. рис. 5, где изображено двухмерное пространство). В других случаях алгоритмы МО решают задачи регрессии или уменьшают число измерений многомерного пространства. Для понимания этих моделей далее может быть применен статистический подход. В оставшейся части данного раздела будут приведены проблемы классификации.

Оптимальная функция дискриминации полностью разделяет классы для обучающей выборки. Можно предположить, что существует точная («верная») функция дискриминации (красная кривая на рис. 5), но на практике с помощью алгоритмов МО определяется лишь приближение истинной функции. Тем не менее, между двумя классами остается разрыв, и не существует однозначного решения для этой проблемы.

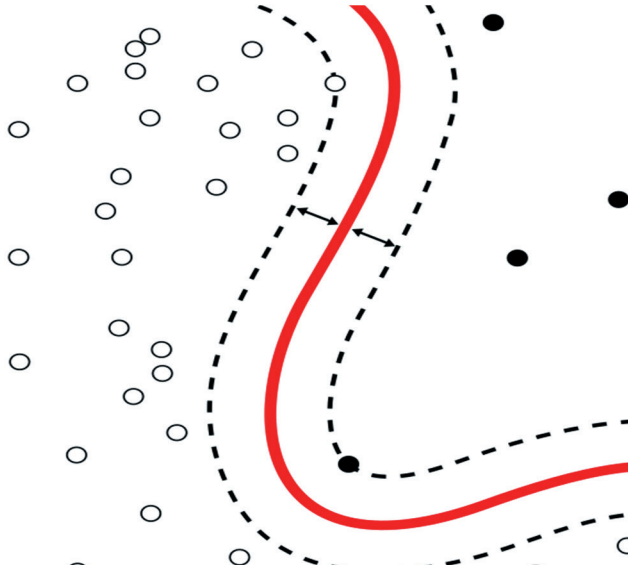

Рис. 5. Дискриминация двух наборов данных при классификации

Искусственные нейронные сети и обобщенная теорема аппроксимации

Наиболее популярным и успешным алгоритмом машинного обучения являются искусственные нейронные сети (ИНС) [4, 11]. Каждая ИНС имеет как минимум два слоя, которые соединены между собой связями, имеющими определенные веса. На рис. 6 приведен простой пример ИНС.

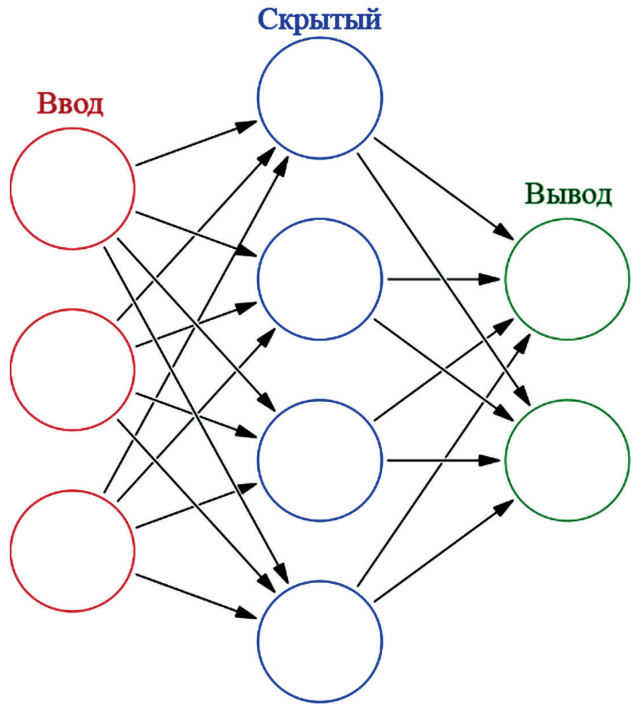

Рис. 6. Искусственная нейронная сеть с двумя слоями

Математическая модель простой ИНС может быть описана следующим образом: вектор входных данных x преобразуется весами v и w, смещениями b и результирующей функцией φ (непостоянной, ограниченной и непрерывной) в два выходных класса:

$$F(x) = \sum_{i=1}^{N} v_i \phi\left(w_i^{\mathrm{T}} x + b_i\right) \qquad (1)$$

Оптимальные веса для определенной функции потерь C, которая формулируется как дополнение к (1), определяются итеративно на основе набора данных для обучения и с использованием численного алгоритма.

При построении более сложных ИНС используются дополнительные скрытые слои (часто называемые

глубокими нейронными сетями), но математическое описание и решение аналогичны тем, что используются при построении простых ИНС.

Из вышеуказанного обзора возникают следующие вопросы:

1. Является ли функция F правильной для четкого различения данных?

2. Является ли она максимально приближенной к истинной функции?

3. Необходимо ли большее количество слоев или более сложные функции?

4. Как подтвердить то, что набор данных для обучения является репрезентативным?

5. Как обнаружить выбросы?

6. Как можно обосновать функцию потерь C?

Если ответы на данные вопросы будут недостаточно всеобъемлющими, то в модели могут возникнуть системные ошибки!

Согласно статье Цибенко [6], для ответа на первый вопрос существует множество так называемых «универсальных теорем приближения», которые демонстрируют сближение функции F с f (истинной функцией), при условии, что φ является ограниченной и непрерывной функцией, а f – непрерывной. Необходимо отметить, что сходимость рассматривается как расчетный метод, а не как стохастическая сходимость.

Это довольно хороший результат, и он также связан с ответами на другие вопросы. Наиболее ограничивающим допущением является непрерывность истинной функции f, что означает, что пространство задач должно быть разделимым с помощью непрерывной функции. Функция φ также должна быть непрерывной, потому что нельзя использовать скачкообразную функцию для принятия решения.

На первый взгляд этот результат удивителен, поскольку он непосредственно относится к ИНС с одним скрытым слоем, но, с другой стороны, результаты довольно очевидны и имеют простое объяснение:

1) F – это некое общее приближение к f. Но очевидно, что такое линейное приближение для непрерывной функции f возможно только в том случае, если имеется достаточно большое количество узлов N. Как показано на рис. 5 (пример классификации), функция f может быть аппроксимирована ступенчатыми линейными функциями.

2) Глубокая ИНС с несколькими скрытыми слоями может быть представлена с помощью одного слоя (с большим количеством узлов N). Предположим, что истинная функция f может быть функцией, представленной многослойной сетью, которая по теореме аппроксимации может быть аппроксимирована однослойной функцией F.

Для надежного применения метода, при ответе на первый вопрос необходимо соблюдать следующие условия:

1) Выбрать однослойную ИНС с достаточно большим количеством узлов N. Количество узлов N можно определить по критерию сходимости согласно расчетному методу.

2) Наиболее сложное предположение, которое необходимо обосновать, состоит в том, что наборы данных могут быть разделены непрерывной функцией. Этот аргумент будет зависеть от типа системных данных и вряд ли может быть общим.

3) Выбрать подходящую функцию потерь C (с обоснованием).

Данные и точность подгонки

Второй вопрос касается качества набора данных для обучения, а также соответствующего правила остановки: когда обучение закончится?

Обеспечение репрезентативности данных означает, что обучение должно происходить в типичной для этого типа системы среде, а среда должна быть такой, чтобы воздействие было типичным для этого типа использования, включая все изменения в среде. Все копии системы (после обучения) должны работать, как минимум, в аналогичной среде и должны быть одинаковыми (см. статью Брабанда, 2018 [2]). Необходимо также учесть проблему обнаружения аномалий (данная проблема касается вопроса 3). Возможно, необходимо ввести правила, связанные с обеспечением безопасности системы, для среды, в которой в которой будет функционировать система.

Следующий вопрос заключается в точности подгонки данных. Как определить точность подгонки набора данных для обучения? Можно ли допустить ошибку в наборе данных для обучения? Как правило, на практике любая неправильная классификация в наборе данных для обучения может привести к высокой доле ошибок при классификации. В качестве примера рассмотрим черную точку на границе линии на рис. 5. Предположим, что оба набора данных разделены истинной (красной) функцией f, как показано на рис. 5. Если данная черная точка классифицирована неправильно, весь набор находящихся рядом с ней точек также будет неверно классифицирован, что приведет к большому количеству ошибок. С другой стороны, данная точка может быть выбросом.

Это означает, что:

1. Либо имеется 100% правильная классификация набора данных для обучения,

2. Либо можно точно рассчитать вероятность возникновения ошибки.

Проблема состоит в том, что невозможно просто посчитать ошибки классификации. Необходимо присвоить им вес в соответствии с их важностью, что может быть затруднительно в больших пространствах и при наличии больших данных.

Кроме того, обучение происходит с использованием статистического подхода, что означает следующее:

- необходимо учитывать доверительные границы;
- производные параметры – это случайные значения с некоторым разбросом;

• последующие решения ИИ также будут случайными с некоторыми ошибками:

- первый тип ошибки: неверное решение, несмотря на то, что входные данные находятся в «правильном» домене;

- второй тип ошибки: входные данные находятся в «неправильном домене», но решение является «верным».

В результате ИИ обладает вероятностью ошибочного решения. Данный фактор необходимо учитывать при выделении части бюджета допустимой интенсивности опасных случайных отказов ИИ (в данном случае: алгоритм).

5. Стандарты функциональной безопасности искусственного интеллекта и возможная процедура оценки

Если ИИ применяется в приложениях, связанных с безопасностью, в действие вступают стандарты функциональной безопасности. Обратимся к базовому стандарту МЭК 61508 [10]. Он представляет собой пример требований стандартов функциональной безопасности. Основная информация содержится в МЭК 61508-3, таблица А.2:

№ 5 – Исправление ошибок методами искусственного интеллекта SIL 2- SIL 4: NR (см. С.3.12);

№ 6 – Динамическая реконфигурация SIL2 – SIL 4: NR (см. С3.13).

В части МЭК 61508-7 приведено определение искусственного интеллекта в рамках стандарта.

С.3.9 Исправление ошибок методами искусственного интеллекта

Для различных каналов связи системы прогнозирование (вычисление тенденций), исправление ошибок, обслуживание и контролирующие действия могут достаточно эффективно поддерживаться системами, основанными на методах искусственного интеллекта (AI). Правила для таких систем могут быть созданы непосредственно из спецификаций и проверены на соответствие. С помощью методов искусственного интеллекта некоторые ошибки общего характера, попадающие в спецификации, для устранения которых уже существуют некоторые правила проектирования и реализации, могут быть исключены, особенно при представлении комбинаций моделей и методов функциональным или описательным способом. Методы выбираются так, чтобы ошибки могли быть устранены и влияние отказов минимизировано для обеспечения требуемой полноты безопасности.

Фактически, МЭК 61508 рассматривает ИИ как средство исправления ошибок, а динамическую реконфигурацию – как реакцию на ошибку в системе управления. Такое применение сделает систему управления непредсказуемой.

Как выполнить требования МЭК 61508, относящиеся к искусственному интеллекту? Утверждение, приведенное в стандарте, связано с утверждением о динамической реконфигурации, что является нежелательным для SIL 2… SIL 4. Если ИИ применяется в самой системе управления, это не будет реакцией на неисправности системы управления, это будет его свойством.

Стандарт функциональной безопасности требует предсказуемой системы. Предсказуемая система означает, что меры, предусмотренные против систематических отказов, позволяют влиянием этих отказов пренебречь. Вероятность возникновения случайных отказов находится на достаточно низком уровне.

Таким образом, поведение системы с ИИ должно быть предсказуемым в статистическом смысле. Стоит учитывать, что предсказуемое поведение означает не детерминированное поведение, а статистически предсказуемое поведение. Это значит, что система ИИ будет способствовать случайным опасным отказам, вызванным случайным поведением самого программного обеспечения. Это является ключевым отличием от обычных систем Э/Э/ПЭ, где программное обеспечение считается детерминированным, а требования относятся только к систематическим ошибкам, благодаря чему выполнение требований к программному обеспечению стандартов функциональной безопасности позволяет снизить вероятность этих ошибок до приемлемого уровня.

Предлагаемый подход к оценке содержит следующие шаги:

• анализ модели;

• выделение части бюджета допустимой интенсивности опасных случайных отказов в системе ИИ, поскольку она демонстрирует вероятностное поведение;

• рассмотрение системы ИИ в качестве нормальной математической модели, но с случайным поведением.

Затем оценка проводится так же, как и обычная оценка безопасности со сложной математической моделью. Авторы не будут приводить всю процедуру оценки безопасности в данной статье. Подробнее о процессе оценки см. Вигтер [17].

Основной частью оценки является проверка модели.

Математическую модель необходимо проверить в отношении следующих аспектов:

• правильность модели с точки зрения физических/химических/математических и других научно обоснованных теорий;

• эквивалентность другим математическим моделям, например, моделям кривых торможения, тепловым моделям и т.д.

Это означает, что теория / модель должна быть открыта эксперту-оценщику. Модели могут быть одного из следующих типов (см., например, Ванг, 2017 [16]):

• нейронная сеть;

• долгая краткосрочная память;

• автокодировщик;

• глубокая машина Больцмана;

• генеративно-состязательная сеть;

• долгая краткосрочная память на основе внимания.

Чем более гибкой является модель, тем сложнее будет ее анализ. В следующем разделе приводится пример того, как подобный анализ может быть выполнен для очень простой модели.

Из-за значительных усилий, необходимых для проверки модели, возникает вопрос, могут ли быть использованы проверенные на практике подходы. Согласно Брабанду и др. [2], это будет означать накопление минимального количества часов безотказной работы (здесь: нет опасных отказов) по следующей схеме:

- $3 \cdot 10^6$ часов безотказной работы для SIL 1;
- $3 \cdot 10^8$ часов безотказной работы для SIL 4.

Практический опыт показывает, что такое количество часов безотказной работы трудно накопить. Как следствие, необходимо провести анализ модели, который остается одной из основных частей оценки безопасности.

6. Академический пример

Целью приведенного в данном разделе примера является не представление модели системы ИИ, а общее описание того, как можно проводить оценку безопасности. Предположим, что система классификации распределяет объекты по двум категориям: «левый» и «правый» – на основе одного действительно-значного параметра. Параметр считается нормально распределенным. Следует обратить внимание на то, что статистически модель полностью определяется этим предположением, которое при использовании на практике должно быть обосновано. Его нельзя воспринимать как данность, поэтому мы называем этот пример теоретическим, ведь в нем предполагается, что нам известна истинная модель.

Два полученных подмножества имеют следующие характеристики:

- «левый» характеризуется нормальным распределением со средним значением m_L и стандартным отклонением σ_L;
- «правый» характеризуется нормальным распределением со средним значением m_R и стандартным отклонением σ_R.

Сначала предположим, что параметры известны.

Затем устанавливается следующее правило классификации:

«левый» если $X \le z$ и «правый» если $X > z$, где z – «правильно» выбранная константа. Теперь могут быть рассчитаны ошибки первого и второго рода.

$$\alpha = 1 - \Phi(z - m_L/\sigma_L) \text{ – ошибка первого рода;} \quad (2)$$

$$\beta = \Phi(z - m_R/\sigma_R) \text{ – ошибка второго рода;} \quad (3)$$

$\Phi(z - m_L/\sigma_L)$ – вероятность правильного отнесения к категории «левый»; (4)

$1 - \Phi(z - m_R/\sigma_R)$ – вероятность правильного отнесения к категории «правый»; (5)

Φ – функция стандартного нормального распределения.

Ошибкой первого рода является вероятность того, что объект будет отнесен к подмножеству «правый», в то время как он принадлежит подмножеству «левый». Ошибкой второго рода является вероятность того, что объект будет отнесен к подмножеству «левый», в то время как он принадлежит подмножеству «правый». Чтобы ошибки были небольшими, параметры σ_R и σ_L должны быть как можно меньше.

Однако существует одна проблема. Параметры m_L, m_R, σ_L и σ_R неизвестны и должны быть получены статистическим образом, т.е. рассчитаны на выборке данных.

Как система обучается? Система обучается на двух выборках для обоих подмножеств: для обучения используются «левая» выборка XL_i, $i = 1, \ldots, n_L$ и «правая» выборка XR_i, $i = 1, \ldots, n_R$.

На основе выборок можно оценить неизвестные точечные параметры:

$$m_R = \frac{1}{n_R}\sum_i XR_i; \quad (6)$$

$$m_L = \frac{1}{n_L}\sum_i XL_i; \quad (7)$$

$$\sigma_R^2 = \frac{1}{(n_R - 1)}\sum_i (XR_i - m_R)^2; \quad (8)$$

$$\sigma_L^2 = \frac{1}{(n_L - 1)}\sum_i (XL_i - m_L)^2. \quad (9)$$

В (6) – (9) суммы рассчитываются по индексу i от 1 до n_L или n_R соответственно.

На следующем шаге вместо точечных оценок, заданных (6) – (9), должны использоваться границы доверительного интервала параметров. Границы доверительного интервала выбираются таким образом, что ошибка, ведущая к неправильной классификации, становится малой, то есть верхние границы для стандартных отклонений и m_R и нижняя граница для m_L. Мы используем однопараметрические границы, а не комбинированные, чтобы упростить вычисления.

Точечные оценки (6) – (9) имеют следующие характеристики:

$(n_L - 1)\sigma_L^2 / \hat{\sigma}_L^2$, где $\hat{\sigma}_L^2$ – дисперсия «левой» генеральной совокупности, имеет распределение хи-квадрат с $n_L - 1$ степенями свободы;

$(n_R - 1)\sigma_R^2 / \hat{\sigma}_R^2$, где $\hat{\sigma}_R^2$ – дисперсия «правой» генеральной совокупности, имеет распределение хи-квадрат с $n_R - 1$ степенями свободы;

$\sqrt{n_L}(m_L - \hat{m}_L)/\hat{\sigma}_L$, где $\hat{m}_L$, $\hat{\sigma}_L$ – соответственно среднее и стандартное отклонение «левой» генеральной совокупности, имеет t-распределение с $n_L - 1$ степенями свободы;

$\sqrt{n_R}(m_R - \hat{m}_R)/\hat{\sigma}_R$, где $\hat{m}_R$, $\hat{\sigma}_R$ – соответственно среднее и стандартное отклонение «правой» генеральной совокупности, имеет t-распределение с $n_R - 1$ степенями свободы.

Наименее благоприятными значениями являются:

верхние границы доверительного интервала дисперсий, т.е.

$$\sqrt{(n_R - 1)/\mathrm{Chi2}(n_R - 1; 1-\gamma)} \cdot \sigma_R, \quad (10)$$

$$\sqrt{(n_L - 1)/\mathrm{Chi2}(n_L - 1; 1-\gamma)} \cdot \sigma_L, \quad (11)$$

где Chi2(n; 1 – γ) – квантиль распределения хи-квадрат с обеспеченностью 1 – γ;

нижняя граница доверительного интервала m_L

$$m_L - \frac{t(n_L - 1, \gamma) \cdot \sigma_L}{\sqrt{n_L}} \quad (12)$$

и верхняя граница доверительного интервала m_R

$$m_R + \frac{t(n_R - 1, \gamma) \cdot \sigma_R}{\sqrt{n_R}}, \quad (13)$$

где $t(n; \gamma)$ – квантиль t-распределения с n степенями свободы и обеспеченностью 1 – γ.

Путем подстановки границ доверительных интервалов (10) – (13) в уравнения (2) – (5) получаем вероятности ошибок.

Если неправильная классификация при ошибке первого типа опасна, из (2) с (6) и (8) получаем вероятность опасного отказа. Однако для учета ошибок, возникающих при применении доверительных интервалов, необходимо использовать значение вероятности ошибки $\alpha + 2\gamma$.

Интерпретация γ как вероятности того, что истинное значение лежит за пределами доверительного интервала, является не частотной, а байесовской с использованием соответствующей априорной вероятности.

Для системы с SIL 1 вероятность отказа по требованию не должна превышать 0,1. Данное значение можно рассматривать как бюджет:

Можно выбрать 0,05 в качестве максимального значения для аппаратных сбоев и 0,05 для алгоритма ИИ. Последнее также можно разделить:

$0{,}05 = \alpha + 2\gamma$,

например, следующим образом:

$\alpha = 0{,}025$, $\gamma = 0{,}0125$.

Для SIL 4 в МЭК 61508 приведено пороговое значение 0,0001 вероятности отказа по требованию.

Читатель может повторить приведенные расчеты. В качестве дальнейшего упражнения можно рассмотреть условия, при которых значения математических ожиданий и стандартных отклонений удовлетворяют требованиям. Этот простой пример показывает, что следует ожидать сложных вычислений. Даже в этом очень простом примере мы столкнулись со сложной математикой.

Как же выйти из этой сложной ситуации?

Существуют два основных варианта:

1. Система ИИ не нуждается в SIL, поскольку ее поведение не имеет критических для безопасности последствий (травм для людей и т.д.).

2. Система ИИ поддерживается достаточно простой системой Э/Э/ПЭ, имеющей необходимый SIL, которая проверяет все опасные решения в соответствии с более простыми алгоритмами и предотвращает опасные реакции.

Эти варианты должны быть подкреплены анализом рисков (см. МЭК 61508).

7. Дальнейшие исследования

Мы признаем, что приведенный пример является достаточно простым и теоретическим, однако мы считаем, что необходимо понять и решить небольшие проблемы, прежде чем подходить к многомерным задачам.

Для рассмотрения чуть более практического примера, разберем следующую задачу: дано множество из n двумерных точек, которые разделены на два подмножества (как на рис. 5, но точки). Модель неизвестна, но количество точек в определенной степени можно контролировать. Известно только то, что задача связана с безопасностью и имеет SIL x. Можно выбрать свой любимый метод классификации, например, ИНС.

При каких допущениях вы сможете представить доказательство безопасности в соответствии с признанным стандартом безопасности, например, МЭК 61508? Можно ли также предоставить адекватное руководство по практической проверке обоснованности этих допущений?

Это может показаться простой проблемой, но она имеет большое значение: если мы не сможем представить доказательство безопасности (при допущениях, которые можно обоснованно проверить на практике), то алгоритмы ИИ (по крайней мере, некоторые классы) не могут быть использованы для приложений, связанных с безопасностью. Но если мы сможем решить эти проблемы при определенных условиях, то используемый подход, возможно, удастся обобщить для больших размерностей.

8. Заключение

В данной статье описан возможный подход к оценке безопасности систем ИИ, однако некоторые вопросы остаются открытыми и могут быть решены только в контексте конкретного приложения.

Уровень полноты безопасности можно определить как для обычной системы Э/Э/ПЭ. Он также должен быть подтвержден анализом опасности и риска. Это необходимо, даже если система не требует SIL.

ИИ можно легко применять в ситуациях, где не возникает критичных для безопасности последствий, что должно быть подтверждено анализом риска. В этом случае не нужно вводить требования к уровню полноты безопасности системы, и оценка безопасности не требуется.

Предложен подход к анализу модели. Проводимый анализ в значительной степени зависит от типа модели. Оценка всегда требует глубокого анализа модели ИИ, что означает, что сам ИИ не может быть проанализирован,

поскольку охватывает множество различных подходов. Чем более гибкой является модель, тем сложнее должен быть ее анализ. При использовании в критичных для безопасности системах удобным подходом может быть ограничение типов моделей для упрощения проектирования и оценки системы ИИ.

Перл и Маккензи [12] подошли к проблеме с той же точки зрения и пришли к выводу, что в ИИ должна быть введена причинно-следственная связь, прежде чем мы сможем полагаться на его выводы. Один из их выводов заключается в том, что необходимо «сформулировать модель процесса, который генерирует данные, или, по крайней мере, некоторые аспекты этого процесса».

Был приведен теоретический пример для демонстрации подхода к анализу конкретного типа модели.

Также представлена задача для дальнейших исследований, решение которой может быть ключевым в использовании алгоритмов ИИ для приложений, связанных с безопасностью. Задача состоит в том, чтобы сформулировать модель процесса генерации данных, которая позволяет проводить анализ безопасности и может быть обоснована для практического применения.

Есть только две возможности использования систем ИИ без необходимости тщательной оценки безопасности: либо иметь систему ИИ, которая не связана с безопасностью, либо иметь другую систему Э/Э/ПЭ, связанную с безопасностью, которая принимает на себя полную ответственность за безопасность.

Библиографический список

1. Anscombe F.J. Graphs in Statistical Analysis // American Statistician. 1973. 27(1). P. 17–21.

2. Braband J., Gall H., Schäbe H. Proven in Use for Software: Assigning an SIL Based on Statistics / Mahboob Q., Zio E., editors. Handbook of RAMS in Railway systems – Theory and Practice. Boca Raton, Taylor and Francis. 2018.

3. Brunette E.S., Flemmer R.C., Flemmer C.L. A review of artificial intelligence // Proc. 4th International Conference on Autonomous Robots and Agents. Feb. Wellington. 2009. P. 385-392.

4. Chen S.H., Jakeman A.J., Norton J.P. Artificial Intelligence techniques: An introduction to their use for modelling environmental systems // Mathematics and Simulation. 2008. Vol. 78. P. 379-400.

5. Corni M. Is Artificial Intelligence Racist? (And Other Concerns). URL : https://towardsdatascience.com/is-artificial-intelligence-racist-and-other-concerns-817fa60d75e9 [accessed October 25, 2018].

6. Cybenko G. Approximations by superpositions of sigmoidal functions // Mathematics of Control, Signals, and Systems. 1989. № 2(4). P. 303–314.

7. Hättasch N., Geisler N. The Deep Learning Hype: Presentation at 36C3. 2019. URL: https://www.youtube.com/watch?v=FomrN5XHQhY.

8. EN 50128. Railway applications – Communication, signalling and processing systems – Software for railway control and protection systems; 2011.

9. EN 50129. Railway applications – Communication, signalling and processing systems – Software for railway control and protection systems; 2018.

10. IEC 61508. Functional safety of electrical/electronic/programmable electronic safety-related systems; 2010.

11. Ivanov A.I., Kuprianov E.N., Tureev S.V. Neural network integration of classical statistical tests for processing small samples of biometrics data // Dependability. 2019. № 19(2). P. 22-27.

12. Pearl J., Mackenzie D. The Book of Why. Penguin Science, 2018.

13. Putzer H. Ein strukturierter Ansatz für funktional sichere KI. Presentation at DKE Funktionale Sicherheit. Erfurt; 2019. (in Ger.)

14. Schäbe H. SIL Apportionment and SIL Allocation. / Mahboob Q., Zio E., editors. Handbook of RAMS in Railway systems – Theory and Practice. Boca Raton, Taylor and Francis. 2018. P. 69-78.

15. Underwriter Laboratories: Standard for Safety for the Evaluation of Autonomous Products. Draft UL 4600; 2019.

16. Wang J., Ma Y., Zhang L., et al. Deep learning for smart manufacturing: Methods and Applications // Journal of Manufacturing Systems. 2017. № 48. P. 144-156.

17. Wigger P. Independent Safety Assessment – Process and Methodology / Mahboob Q., Zio E., editors. Handbook of RAMS in Railway systems – Theory and Practice. Boca Raton, Taylor and Francis. 2018. P. 475-485.

Сведения об авторах

Йенс Брабанд – доктор естествознания, главный эксперт по RAMSS at Siemens Mobility GmbH, профессор Технического Университета, Брауншвейг, Германия, e-mail: jens.braband@siemens.com

Шебе Хендрик – доктор физико-математических наук, заведующий отделом анализа рисков и опасностей, TÜV Rheinland InterTraffic, Кельн, Германия, e-mail: schaebe@de.tuv.com

Вклад автора в статью

Вклад авторов заключается в анализе систем с искусственным интеллектом как статистических моделей, анализ подхода к оценке безопасности таких систем и рассмотрении примера. Вклад авторов равный.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.