УДК 519.25: 004.891.3 DOI: 10.26102/2310-6018/2020.28.1.020

Математическая модель обнаружения аномальных наблюдений с использованием анализа чувствительности нейронной сети

Р.В. Щеглеватых[1], А.С. Сысоев[2]
[1] *Федеральный фонд обязательного медицинского страхования*
Москва, Россия
[2] *Липецкий государственный технический университет*
Липецк, Россия

Резюме: Переход к цифровизации в различных сферах экономической и социальной деятельности сопровождается возникновением больших массивов данных, обрабатывая которые, необходимо выявлять определенные зависимости, строить модели процессов и систем. Актуальной является задача поиска аномальных значений в больших массивах данных. Существующие алгоритмы выявления аномалий основываются на использовании различных подходов и имеют свои преимущества и недостатки. Однако базовые схемы работы всех методов схожи – на начальном этапе происходит разделение данных на типичные для системы или процесса и те, которые не вписываются в общую картину, затем происходят структурная и параметрическая идентификация модели, на заключительном этапе обученная модель используется для разделения данных. Для повышения точности работы алгоритмов возможны их модификации, учитывающие структуру данных или позволяющие комбинировать разнородные математические модели. В статье приводится описание комбинированного подхода к построению системы обнаружения аномальных реализаций на основе алгоритма изолирующего леса и последовательного применения нейросетевого классификатора. Для снижения размерности входного вектора нейросетевой модели синтезирован и описан подход к анализу чувствительности по факторам нейросетевой модели, основанный на применении анализа конечных изменений. Приведен численный пример, показывающий адекватность применимости предлагаемого подхода к анализу данных.

Ключевые слова: математическая модель, аномальные значения, анализ чувствительности, нейросетевые модели.

Для цитирования: Щеглеватых Р.В., Сысоев А.С. Математическая модель обнаружения аномальных наблюдений с использованием анализа чувствительности нейронной сети. *Моделирование, оптимизация и информационные технологии.* 2020;8(1). Доступно по: https://moit.vivt.ru/wp-content/uploads/2020/02/ScheglevatychSysoev_1_20_1.pdf DOI: 10.26102/2310-6018/2020.28.1.020

Mathematical model to detect anomalies using Sensitivity Analysis applying to neural network

R.V. Scheglevatych[1], A.S. Sysoev[2]
[1] *Federal Compulsory Medical Insurance Fund*
Moscow, Russia
[2] *Lipetsk State Technical University*
Lipetsk, Russia

Abstract: The transition to the digitalization in various spheres of economic and social activity is accompanied by the emergence of large amounts of data, processing which it is necessary to identify certain dependencies and build models of processes or systems. The task to identify anomaly values in dig data sets is relevant. Existing algorithms to detect anomalies are based on different approaches and have their own advantages and disadvantages. However basic schemes of all methods are similar and use at the initial stage the separation of data in a typical for system or process and those that are not, then follow structural and parametric identification of the model, and at the final stage the trained model is used to separate the data. To improve the accuracy of algorithms, they can be modified to take into account the data structure or to combine heterogeneous mathematical models. The paper describes a combined approach to build the system for detecting anomalies based on the Isolation Forest algorithm and sequential application of a neural network classifier. To reduce the dimension of neural network input vector, the approach to Sensitivity Analysis based on applying Analysis of Finite Fluctuations to the neural network model is synthesized and described. It is presented the numerical example that shows the adequacy of the proposed approach to data analysis.

Keywords: mathematical model, anomalies, sensitivity analysis, neural-network models.

For citation: Scheglevatych R.V., Sysoev A.S. Mathematical model to detect anomalies using Sensitivity Analysis applying to neural network. *Modeling, Optimization and Information Technology*. 2020;8(1). Available from: https://moit.vivt.ru/wp-content/uploads/2020/02/ScheglevatychSysoev_1_20_1.pdf DOI: 10.26102/2310-6018/2020.28.1.020 (In Russ).

Введение

Переход к цифровизации в различных сферах экономической и социальной деятельности сопровождается возникновением больших массивов данных, обрабатывая которые, необходимо выявлять определенные зависимости, строить модели процессов и систем. Актуальной является задача поиска аномальных значений в больших массивах данных. Преждевременное определение аномального значения способно предотвратить сбои в работе сложных технических систем, техногенные аварии, выявить нарушения в фиксации экономических и социальных показателей и тем самым привести к экономии бюджетных средств.

Выявление аномалий относится к проблеме нахождения данных, не соответствующих некоторому ожидаемому поведению процесса или показателю, возникающему в системе. При построении систем обнаружения аномальных наблюдений большое внимание необходимо уделять модели, лежащей в основе системы. Исследование посвящено построению системы обнаружения аномальных значений фиксируемого показателя на основе комбинации алгоритма изолирующего леса для оценки показателя аномальности наблюдения и последующего применения нейросетевого классификатора. Для оценки того, какие факторы должны быть переданы на вход нейросетевого классификатора, был синтезирован подход к анализу чувствительности нейросетевой модели, основанный на применении анализа конечных изменений.

Цель представленного исследования – построить математическую модель обнаружения аномальных наблюдений и подход к определению влияния, которое факторы оказывают на отклик рассматриваемой системы.

Комбинированный подход к обнаружению аномальных наблюдений

Виды аномальных наблюдений. Исходя из ситуаций, в которых возникают аномалии, такие наблюдения можно разделить на три группы: изолированные (единичное наблюдение может быть отнесено к этой группе в случае, если оно сильно отличается от имеющегося массива имеющихся данных), контекстуальные (наблюдение является принципиально возможным, но нереализуемым в некоторых сложившихся условиях; для идентификации такого типа аномалий необходимо принимать во внимание все существующие в рассматриваемой системе ограничения на области допустимых значений факторов), коллективные (несколько последовательных наблюдений образуют аномальное множество по отношению к остальным данным; по отдельности же эти наблюдения могут и не быть аномалиями, но существование их в некоторой последовательности говорит о наличии аномального состояния процесса или системы) [1].

Схема обнаружения аномальных наблюдений. Большинство существующих работ, посвященных поиску и исследованию структуры аномальных наблюдений, предлагают два типа подходов к выявлению (детектированию) аномальных наблюдений.

Первый предполагает присвоение по результатам анализа каждому из наблюдений метки «нормальное» и «аномальное» наблюдение. Второй тип подходов основан на расчете некоторого определенного показателя для каждого из наблюдений с целью отнесения его к одной из указанных выше групп.

В общем схему работы любого алгоритма обнаружения выбросов можно представить ка следующую последовательность действий:

- параметризация (подготовка анализируемых данных с целью дальнейшего поиска аномалий; выявляются подозрительные наблюдения);
- выбор типа и конструирование модели для обнаружения выбросов (модель может быть построена для обнаружения данных, описывающих нормальное поведение системы или аномальные ситуации);
- обнаружение аномалий с помощью построенной модели (полученные с помощью построенной модели показатели для каждого наблюдения сравниваются с выявленными на этапе параметризации подозрительными наблюдениями и, на основании некоторого порога (меры схожести) происходит отнесение данных к категории «нормальных» или «аномальных»).

Подходы к обнаружению аномалий. Одной из групп подходов к обнаружению аномалий являются алгоритмы, основанные на применении методов классификации. Указанная группа методов предполагает на основании некоторой построенной с помощью обучающей выборки модели маркировку новых наблюдений для отнесения к категории «нормальных» или «аномальных» наблюдений. В качестве инструментов для классификации могут быть использованы различные структуры: деревья решения, модели нечеткой логики, наивные байесовские модели, генетические алгоритмы, нейронные сети, опорные вектора и т.п. При использовании деревьев решения структуру дерева формируют «листья» (концевые вершины) и «ветки» (ребра). Ребра содержат атрибуты, от которых зависит целевая переменная, а концевые вершины – значения целевой переменной, остальные узлы (вершины) – атрибуты, по которым различаются случаи. Для классификации нового случая, необходимо пройти от вершины дерева до листа и выдать соответствующее значение. Для построения нечетких моделей необходимо на основании некоторых оценок разделить массив наблюдаемых данных. В качестве таких оценок могут выступать различные статистические метрики. Для анализируемых данных строится база нечетких правил логического вывода для классификации новых наблюдений. В реальных системах могут существовать статистические связи между переменными, которые достаточно сложно описать до построения модели. В таком случае возможно использование вероятностной графовой модели, называемой наивной байесовской сетью, которая способна учитывать структурные связи между переменными случайной природы. При таком подходе предполагается строить направленный ациклический граф, являющийся обобщением дерева. В построенном графе каждая вершина представляет собой переменную системы, а дуга отражает степень влияния одной переменной на другую. По сравнению с деревьями решений, наивные байесовский модели проигрывают в точности, однако, заметно выигрывают с точки зрения скорости вычислений. Поэтому, оправданным является применение такого подхода к данным большого объема. Основанные на аналогах биологических процессов наследования, селекции, мутации и кроссовера, генетические алгоритмы широко применяются для решения задач оптимизации. Известно применение генетического алгоритма в задаче обнаружения вторжения в сеть для построения базы правил классификации из общей базы служебных данных сети. Преимуществами использования этого класса эволюционных алгоритмов является устойчивость по отношению к шуму и возможность самообучения. Нейронная сеть представляет собой набор объектов (нейронов), разделенных на группы (слои), соединенных между собой. Биологическая модель, положенная в основу такой структуры – сеть нервных клеток живого организма. Известны применения нейронной сети при решении задач классификации (с учителем и без). Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора. В ходе реализации происходит перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Для улучшения результатов классификации при решении задачи обнаружения аномальных наблюдений были предложены и комбинированные методы, сочетающие использование нескольких алгоритмов. Среди таких комбинаций можно выделить каскадные техники классификации с учителем (сочетание наивных байесовских моделей и деревьев решений, деревьев решений и метода опорных векторов) и комбинации классификационных схем с учителем и без учителя (например, сочетание метода опорных векторов и классификации методом k-средних).

Отмеченные выше способы нахождения аномальных наблюдений предполагают, что в результате анализа будет построена модель, описывающая профиль «нормального» наблюдения. Однако существует и принципиально отличный подход, основанный на построении модели, определяющей не «нормальное» значение и отвергающей все, но попадающие под такое понимание, а на построении модели, выявляющей значения, отличные от всех, типичных для показателей рассматриваемой системы или процесса. Структура, положенная в основу работы такого метода, - изолирующий лес. Изоляция означает отделение одной группы наблюдений от другой. Чтобы применить такую идею для каждого наблюдения, необходимо вычислить некоторую меру восприимчивости, определяющую порог разделения. Естественные структуры, разделяющие данные, - это случайно сгенерированные двоичные деревья, экземпляры которых рекурсивно разделены [2, 3]. Изолирующий лес строится по принципу Монте-Карло. В деревьях для подозрительных (аномальных) наблюдений формируются выделяющиеся более короткие пути, поскольку (а) в областях с аномалиями меньшее количество аномалий приводит к меньшему числу секций и (б) экземпляры с различимыми значениями атрибутов с большей вероятностью будут разделены в начале процесса секционирования. Поэтому, когда случайный лес в совокупности создает более короткие длины путей для некоторых конкретных точек, они с высокой вероятностью будут аномалиями.

Такой подход имеет ряд преимуществ, в том числе и способность распознавать аномалии различной природы (изолированные и коллективные). С точки зрения вычислительной сложности он более эффективен, чем большинство других алгоритмов. Он не требует значительных затрат памяти и определения метрики или некоторой априорной информации об анализируемом объекте, устойчив к увеличению размерности.

Для выявления аномалий при рассматриваемом подходе необходимо применять двухэтапный процесс. На первом этапе необходимо обучить выбранную модель (задать параметры модели: указать количество деревьев в ансамбле и размер подстановки). Второй этап посвящен анализу полученной модели для получения оценки для каждого наблюдения, с целью отнесения / не отнесения его к категории аномальных.

Этап обучения – это процесс рекурсивного секционирования, на котором каждое дерево строится с использованием подвыборки X', случайно полученной без замены из исходной выборки наблюдений X. Основной шаг алгоритма принимает следующие входные параметры: X – набор начальных наблюдений, t – количество деревьев, используемых для построения модели, и ψ – размер подвыборки. Для каждого дерева от 1 до t случайным образом выбирается новая подвыборка X' из X с учетом ψ и добавляется к текущему лесу. Здесь с помощью параметра ψ возможно управлять размером обучающих данных, он был эмпирически подобран в исследовании [2] (исходя из предположения, что аномалий “мало” и они “отличные”) – это значение 256. При построении каждого дерева выполняется следующий процесс. В качестве входных данных используется случайная подвыборка X' в качестве выходной структуры получается изолирующее дерево $iTree$. В случае, если X' нельзя разделить на два подмножества, то $iTree$ достигнута концевая вершина; в противном случае, пусть Q – список атрибутов в X', следует случайным образом выбрать атрибут q из Q, а также случайным образом выбрать точку разделения p между минимальным и максимальным значениями выбранного атрибута q. Получены два подмножества X_l и X_r как отфильтрованные на основе P пороговых входных значений, которые затем образуют левый и правый внутренние узлы. Следует отметить, что временная сложность этой стадии равна $O(t\psi^2)$, а пространственная – $O(t\psi)$.

На втором этапе, этапе оценки (входные данные: x – наблюдение, T – структура изолирующего леса $iTree$, $hlim$ – порог отнесения наблюдения к разряду аномалий, e – текущая длина пути, которая равна 0 на первом шаге) длина пути $h(x)$ получается путем подсчета количества ребер e от корневого узла к концевой вершине, поскольку наблюдение x проходит через $iTree$. Когда обход достигает предопределенного предела высоты $hlim$, то возвращается значение $e + c$ (c – некоторая поправка). Эта корректировка позволяет оценить среднюю длину пути случайного поддерева, которое может быть построено с использованием данных размера, превышающего предельный порог. Когда значение $h(x)$ получено для каждого дерева ансамбля, вычисляется оценка аномалии. Временная сложность процесса оценки равна $O(nt\psi)$, где n – размер выборки данных.

Задача обнаружения аномалий заключается в определении в построенном дереве некоторой характеристики для каждого наблюдения, отражающей степень его аномальности. Поэтому, чтобы определить, является ли наблюдение аномальным, необходимо отсортировать данные о длинах путей $h(x)$ в дереве или о степени аномалии. Поскольку случайный лес имеет эквивалентную двоичному дереву поиска структуру, можно применить ту же процедуру для оценки среднего пути $h(x)$ для внешнего узла, что и в бинарном лесе. В [3] показано, что средняя длина пути безуспешных поисков

$$c(\psi) = 2H(\psi - 1) - 2(\psi - 1) / n \text{ для } \psi > 2,$$
$$c(\psi) = 1 \text{ для } \psi = 2 \text{ и}$$
$$c(\psi) = 0 \text{ иначе.}$$

Здесь $H(i) = \ln(i) + 0.5772156649$ (константа Эйлера) – гармоническое число.

Степень аномальности s для наблюдения x определяется как

$$s(x,\psi) = 2^{-E(h(\psi))\,/\,c(\psi)},$$

где $E(h(x))$ – средняя длина пути $h(x)$ в построенном изолирующем лесе.

Следующие условия определяют три специальных значения оценки аномальности наблюдений:

$$\text{если } E(h(x)) \to 0,\ s \to 1;$$
$$\text{если } E(h(x)) \to \psi - 1,\ s \to 0; \text{ и}$$
$$\text{если } E(h(x)) \to c(\psi),\ s \to 0{,}5.$$

Таким образом, когда $s \to 1$, то наблюдение x является аномалией, а когда $s <<$ 0,5, то x является нормальным наблюдением. Если для всех наблюдений значение $s \approx 0.5$, то массив данных X не имеет аномалий.

Описанный метод имеет много преимуществ и хорошо обнаруживает аномалии, в частности он чувствителен к возникновению контекстных аномалий, которые могут быть интерпретированы как технические ошибки фиксации данных или их намеренное искажение. Чтобы обеспечить такое обнаружение, необходимо построить массивы «нормальных» и аномальных реализаций, которые впоследствии будут проанализированы контролирующими специалистами.

Для повышения качества обнаружения аномальных значений предлагается использовать изолирующий лес в качестве первого шага для фильтрации данных (сформировать группы «нормальных» и аномальных наблюдений), затем проанализировать все аномалии и выделить среди них контекстуальные аномалии (принципиально возможные наблюдения, но не типичные по сравнению с ближайшими данными), а затем, применяя нейросетевой классификатор [16–18], построить модель, способную находить выбросы для новых исходных данных:

$$Y^{(n)} = \Psi^{(n)}\Psi^{(n\text{-}1)}...\Psi^{(1)}X\ ,$$

где $Y^{(n)}$ – выход n-слойной нейронной сети (значение, характеризующее принадлежность наблюдения к множеству выбросов),

X – вектор входных факторов,

$\Psi^{(1)}, \Psi^{(2)}, ..., \Psi^{(n)}$ - функции активации слоев нейронной сети.

Однако использование такого подхода требует значительных вычислительных ресурсов, особенно с увеличением числа факторов модели (так как возрастает и число параметров – весов нейронов). Далее приводится подход к анализу чувствительности по факторам нейросетевой модели.

Анализ чувствительности по факторам нейросетевых моделей

Классическим подходом при рассмотрении чувствительности систем является нахождение чувствительности по параметром изучаемой системы [4], однако, существует и направление анализа чувствительности, предполагающее использование в качестве оцениваемых параметров системы ее факторы [5]. Далее приводится подход через призму такой постановки задачи анализа чувствительности модели (показателя системы).

Постановка задачи анализа чувствительности модели. Пусть $X = (x_1,...,x_n) \in R^d$ - вектор факторов системы (ее переменные, входы), $G(\cdot)$ - оператор системы (функция), $Y \in R$ - выход системы (показатель, значение функции), тогда систему можно представить в виде $Y = G(X)$. Необходимо определить, какие влияние каждый из факторов системы оказывает на ее выход. Необходимо заметить, что в качестве оператора G выступать любая модель, связывающая значения выхода с соответствующими значениями входов (например, дифференциальное уравнение, система дифференциальных уравнений, нейронная сеть, программный код и т.п.).

Классификация методов анализа чувствительности по факторам. Существующие методы анализа чувствительности систем можно сгруппировать в пять больших классов [6]. Первое семейство подходов [7] основано на изучении графического попарного представления связи выхода системы и ее входов. Например, линейная связь может быть охарактеризована (измерена) коэффициентом корреляции Пирсона (для переменных числового характера) и ранговым коэффициентом корреляции Спирмена (в случае использования ранговых измерений переменных). Например, при исследовании линейной регрессионной модели оценки, показывающие влияние факторов x_i на Y, - это стандартные регрессионные коэффициенты $SRC_i = \beta_i \sqrt{\frac{\sigma^2_{x_i}}{\sigma^2_Y}}$. В таком случае значения SRC_i^2 показывают разброс вариации Y, объясненный вариацией переменной x_i.

Второе семейство подходов [8, 9] построено на использовании производных. Такое предположение является естественным в силу природы производной функции – она описывает тенденцию увеличения значения функции при изменении значения аргументов в некоторой точке. Описываемый подход состоит в оценке чувствительности G при изменении X, т.е. $\frac{dG}{dX}(X)$, где $G: R^d \to R$ - дифференцируемая функция с непрерывными производными и градиент:

$$\nabla_X G(X) = \left(\frac{\partial G}{\partial x_1}(X),...,\frac{\partial G}{\partial x_d}(X) \right)^T.$$

Простейших подход в таком случае – зафиксировать все переменные модели, кроме изучаемой, и оценить ее влияние на изменение выхода модели. Более сложные подходы предполагают построение численных аппроксимаций градиента и их дальнейшее исследование. Однако указанный подход имеет два крупных недостатка: значительное увеличение вычислительных затрат с возрастанием количества факторов модели и критичный выбор приращения факторов Δx_i. Но вместе с этим эти недостатки дают перспективы применения указанного семейства методов для высокой размерности вектора входов при хорошем выборе приращений факторов. Тогда проблема сводится к построению такой эффективной схемы аппроксимации.

Третье семейство подходов [10-12] основывается на скрининге факторов системы. Качественный выбор факторов необходим для построение точной (адекватной, качественной) модели системы или процесса. Проблемой использования таких подходов – это необходимость изучения большого числа факторов, для которых доступно небольшое число реализаций. Фундаментальным принципом успеха при решении этой проблемы является предположение, что лишь небольшое число входных переменных оказывают существенное влияние на выход. Подавляющее число скрининговых процедур основаны на идентификации и использовании суррогатных моделей, аппроксимирующих выход системы *Y*(*X*). Например, во многих прикладных задачах успешно применялся класс моделей

$$Y(X) = h^T(X)\beta + \varepsilon(X),$$

где h – вектор размера $p \times 1$, составленный из определенных функций от *X*,

$\beta = \left(\beta_0, ..., \beta_{p\text{-}1}\right)^T$ - вектор неизвестных параметров модели,

$\varepsilon(X) \sim N\left(0, \sigma^2\right)$ - случайная составляющая.

Важным практическим вопросом при применении такого подхода является контроль, с которым выбранная суррогатная модель аппроксимирует изучаемую систему, т.е. необходим корректный выбор функций, формирующих вектор *h*(*X*). В качестве суррогатных моделей наиболее часто используются линейные регрессионные модели и модели гауссовских процессов.

Четвертое семейство подходов [13] ориентировано на построении составных показателей, агрегирующих некоторое количество факторов с определенными весами. В качестве составного показателя *y*, который может агрегировать *d* факторов x_i, часто используется взвешенное среднее

$$y_j = \sum_{i=1}^{d} w_i x_{ji}, \; j = 1, ..., n,$$

где x_{ji} - нормализованные значения показателя *j* и w_i - значения весов факторов x_i.

В [5] предложено использовать следующий подход к оценки влияния факторов на отклик системы. Пусть математическое ожидание и дисперсия отклика *y* равны $E(y) = \mu_y$ и $V(y) = \sigma_y^2$ соответственно. Аналогично для каждого фактора x_i имеем $\left\{\mu_{x_i}\right\}_{i=1}^{d}$ и $\left\{\sigma_{x_i}^2\right\}_{i=1}^{d}$.

Пусть $y_j = f_i(x_{ji}) + \varepsilon_j$, где $f_i(x_{ji})$ - подходящая (в общем случае нелинейная) функция, описывающая математическое ожидание $E(y_j \mid x_i)$ показателя y_j при определенном значении x_i, ε_j - случайная составляющая. Тогда

$$\operatorname{cov}(y, x_i) = E\left[(y - \mu_y)(x_i - \mu_{x_i})\right], \; R_i = \frac{\operatorname{cov}(y, x_i)}{\sigma_y \sigma_{x_i}}.$$

В случае линейной регрессионной модели $R_i^2 = r^2(y, x_i)$ - коэффициент детерминации. В случае наличия нелинейной связи между *y* и x_i является смещенной.

Исследование [5] предлагает использование корреляционных отношений S_i, известных также как индексы чувствительности первого порядка. Такие индексы могут быть интерпретированы как ожидаемое снижение дисперсии составного показателя при фиксировании некоторого выбранного фактора:

$$S_i = \frac{V_{x_i}\left(E_{x\sim i}\left(y \mid x_i\right)\right)}{V(y)},$$

где $x \sim i$ - вектор, содержащий все факторы системы $\left(x_1, ..., x_d\right)$, за исключением фактора x_i,

$E_{x\sim i}\left(y \mid x_i\right)$ - условное математическое ожидание показателя *y* для заданного значения *x*.

Поскольку $E\left(y \mid x_i\right)$ есть функция, зависящая от x, то, обозначив $m_j = \hat{f}_i(x_{ij})$, получим $S_i = \frac{\sum_{j=1}^{n}\left(m_j - \bar{m}\right)^2}{\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2}$. Здесь $\bar{m} = n^{-1}\sum_{j=1}^{n} m_j$, $m_j = \hat{m}(x_{ji})$, $\hat{m}(\cdot) = f_i(\cdot)$.

Очевидно, что в случае линейной регрессионной модели $R_i^2 = S_i$. Коэффициенты S_i получили названия коэффициентов Соболя. Подход к оценке коэффициентов Соболя состоит в декомпозиции функции, описывающей модель. Такой подход лежит в основе функционального дискриминантного анализа (FANOVA).

Пятое семейство подходов основано на оценке дисперсии модели. Один из наиболее развитых и представительных методов в этой категории – алгоритм Гарсона [14], используемый для оценки чувствительности нейросетевой модели. Этот подход основан на изучении весовых коэффициентов построенной нейросетевой модели.

Например, для нейронной сети классической структуры с одним скрытым слоем, коэффициенты чувствительности факторов могут быть вычислены как

$$S_k^p(i) = \frac{\sum_{j=1}^{n}\left(w_{ij} \cdot v_{jk} / \sum_{i=1}^{n} w_{ij}\right)}{\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\left(w_{ij} \cdot v_{jk} / \sum_{i=1}^{n} w_{ij}\right)\right)},$$

где i, j и k – индексы весовых коэффициентов входного, скрытого и выходного слоев соответственно.

Анализ конечных изменений. Анализ конечных изменений (АКИ) может быть описан как подход к анализу сложных систем различной структуры с целью построения зависимости, связывающей конечные изменения показателя (функции) с конечными изменениями факторов (переменных). Впервые этот подход был представлен в [15] как логичное расширение экономического факторного анализа и затем нашел применение в различных прикладных исследованиях [16].

Обозначим изменение некоторой величины (фактора) x через $\mu(x)$. Естественной формой такого показателя является абсолютное приращение $\mu(x) = \Delta x = b - a$ при начальном значении фактора $x^{(0)} = a$ и его конечном значении $x^{(1)} = b$.

Основная задача анализа конечных изменений формулируется следующим образом: пусть задана зависимость

$$y = f(X) = f(x_1, ..., x_d),\ \ x \in R^d, \tag{1}$$

описывающая связь выхода системы y и ее входов x_i, $i = 1, ..., n$. Необходимо трансформировать модель (1), чтобы она приняла вид

$$\mu(y) = \varphi(\mu(x_1), ..., \mu(x_d)), \tag{2}$$

связывающий конечные изменения ее входов и выхода.

Стоит отметить, что во многих практических приложениях конечные изменения, отмеченные выше, предполагаются малыми.

Для случая малых конечных приращений из математического анализа известна теорема, позволяющая преобразовать модель (1) к виду (2). Это теорема Лагранжа о средней точке (промежуточной точке, формула конечных приращений). Для случая функции многих переменных, определенной и непрерывной на своей области определения и имеющей на ней частные производные, она формулируется следующим образом:

$$\Delta y = \sum_{i=1}^{d} \frac{\partial f(X^{(m)})}{\partial x_i} \cdot \Delta x_i, \tag{3}$$
$$X^{(m)} = \left(x_1^{(m)}, ..., x_d^{(m)}\right), \quad x_i^{(m)} = x_i^{(0)} + \alpha \cdot \Delta x_i,$$
$$i = 1, ..., d, \quad 0 < \alpha < 1.$$

Промежуточная точка определяется значением параметра α.

АКИ может быть применен для исследования сложных систем с целью выявления наиболее значимых факторов рассматриваемой системы (задача сокращения размерности). В частности, рассматривается проблема нахождения аномальных наблюдений с применением нейросетевого классификатора, число входов которого необходимо разумно сократить.

Пусть задана нейронная сеть, содержащая n скрытых слоев, которая описывает поведение технической, социально-экономической системы или технологического процесса в виде $Y^{(n)} = \Psi^{(n)}\Psi^{(n-1)}...\Psi^{(1)}X$, где $X = \left(x_1, ..., x_d\right)^T$.

В текущий момент времени начальное состояние факторов системы имеет вид $X^{(0)} = \left(x_1^{(0)}, ..., x_d^{(0)}\right)^T$ и выход системы $Y_0^{(n)} = \Psi^{(n)}\Psi^{(n-1)}...\Psi^{(1)}\left(x_1^{(0)}, ..., x_d^{(0)}\right)^T$. В следующий момент фиксации факторы системы претерпели изменения и описываются как $X^{(1)} = \left(x_1^{(1)}, ..., x_d^{(1)}\right)^T$, выход системы $Y_1^{(n)} = \Psi^{(n)}\Psi^{(n-1)}...\Psi^{(1)}\left(x_1^{(1)}, ..., x_d^{(1)}\right)^T$.

Таким образом, приращение выхода системы может быть определено, с одной стороны, как разница нового и предыдущего значений выходов и, с другой стороны, по теореме Лагранжа, т.е. может быть составлено и решено относительно параметра α следующее уравнение:

$$Y_1^{(n)} - Y_0^{(n)} = \sum_{i=1}^{d} \frac{\partial Y^{(n)}}{\partial x_i}\left(..., x_i^{(0)} + \alpha \cdot \Delta x_i, ...\right) \cdot \Delta x_i,$$

что позволит оценить так называемые факторные нагрузки A_{x_i} и получить модель вида

$$\Delta Y^{(n)} = \sum_{i=1}^{d} \frac{\partial Y^{(n)}}{\partial x_i}\left(..., x_i^{(0)} + \alpha \cdot \Delta x_i, ...\right) \cdot \Delta x_i =$$
$$= A_{x_1}\Delta x_1 + ... + A_{x_d}\Delta x_d.$$

Указанная процедура может быть повторена $d-1$ раз, численные результаты анализа (факторные нагрузки) могут быть усреднены и получены оценки влиятельности факторов рассматриваемой системы, что позволит сократить числа факторов системы.

Численный эксперимент. Обсуждение полученных результатов

Эксперимент состоял из двух частей. На первом этапе было определено, насколько влиятельными являются факторы системы, на втором этапе – выявлены аномальные наблюдения в массиве данных. Для реализации всех приведенных в статье алгоритмов было использовано свободное программное обеспечение R.

В качестве исходных данных использован набор данных neuraldat из пакета NeuralNetTools. Этот набор данных включает 3 фактора и один выход. Стоит заметить, что рассматривался лишь вопрос моделирования зависимости $y = f(x_1, x_2, x_3)$ без изучения прогнозных свойств полученной модели. Выборка для анализа содержала 2000 реализаций. Была использована нейросетевая модель, имеющая структуру

$$y_k = \psi_1\left(b_0 + \sum_{j=1}^{3} w_j \psi_2\left(b_1 + \sum_{m=1}^{3} w_{jm} x_m\right)\right),$$

где y_k - модельные значения,

$X = \{x_1, x_2, x_3\}$ - вектор факторов системы,

w_j и w_{jm} - весовые коэффициенты,

b_0 и b_1 - свободные коэффициенты нейронов выходного и скрытого слоев соответственно;

$\psi_1(net) = \psi_2(net) = 1/(1+\exp(-net))$ - логистические функции активации.

Было протестировано три подхода (коэффициенты чувствительности Соболя, использование алгоритма Гарсона и предлагаемый подход, основанный на применении АКИ). Таблица 1 показывает сравнение результатов оценок важности рассматриваемых факторов. Так как тестируемые подходы имеют выходные значения различных форматов, все результаты были нормированы в своих шкалах для сопоставления результатов (и представлены в процентном соотношении). Стоит также отметить, что при применении АКИ было получено 1999 оценок (так как для исходной выборки существует 1999 конечных приращений), медианные значения которых были использованы впоследствии в качестве меры чувствительности модели по факторам.

Таблица 1. Сравнение оценок чувствительности (важность факторов в %)

Подход	**Факторы**		
	x_1	x_2	x_3
Коэффициенты чувствительности Соболя	28,97	47,74	23,29
Алгоритм Гарсона	30,33	46,92	22,75
Анализ конечных изменений	30,96	44,68	24,36

Анализ таблицы 1 показывает, что все сравниваемые подходы дают сопоставимые результаты с небольшими отклонениями, что доказывает, что применение АКИ не является противоречивым в данной проблеме. Но, по сравнению с использованием коэффициентов Соболя, предлагаемый подход не использует аппроксимацию для моделирования статистических параметров исследуемой модели, а по сравнению с алгоритмом Гарсона, АКИ оперирует и параметрами и факторами модели.

Вторая часть исследования заключалась в числа факторов модели. В данном случае был исключен только один фактор x_3. Затем проводилось нейросетевое моделирование зависимости того, является ли рассматриваемая реализация аномалией от значений факторов модели в этой реализации. Используемая нейросетевая модель показала высокие результаты точности, обнаружив на тестовой выборке 94% детектированных аномальных наблюдений.

Заключение

В работе представляет подход к построению системы обнаружения аномальных наблюдений, основанный на применении алгоритма случайного леса для оценки показателя аномальности каждого наблюдения с последующим экспертным анализом полученной информации и построением нейросетевого классификатора. Ввиду возможного большого количества входных факторов для такой системы необходимо был синтезирован подход к чувствительности по факторам применяемой нейросетевой модели. Предлагаемый подход может быть использован для обнаружения аномальных наблюдений (выбросов, контекстных аномалий) в данных об оказании медицинской помощи населению, что упростит процедуру контроля качества этого процесса [17,18].

ЛИТЕРАТУРА

1. Parmar J. D., Patel J. T. Anomaly Detection in Data Mining: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*. (2017);7(4):32-40.
2. Liu F. T., Ting K. M., Zhou Z. H. Isolation forest. *IEEE International Conference on Data Mining, ICDM* (2008).
3. Liu F. T., Ting K. M., Zhou Z. H. Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data*. (2012).
4. Розенвассер Е. Н., Юсупов Р. М. *Чувствительность систем управления*. Москва: Наука. (1981).
5. Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M., Tarantola S. *Global Sensitivity Analysis – The Primer*. John Wiley & Sons, Hoboken (2008).
6. Saltelli A., Tarantola S., Campolongo F. Sensitivity analysis as an ingredient of modeling. *Stat. Sci*. (2000); 15(4):377–395.
7. Kurowicka D., Cooke R. *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley. (2006).
8. Cacuci D. G. *Sensitivity and Uncertainty Analysis Theory*. Chapman and Hall/CRC. (2005).
9. Hamby D. M. A review of techniques for parameter sensitivity analysis of environmental models. *Environ. Monit. Assess*. (1994); 32(2):135–154.
10. Box G. E. P., Meyer R. D. An analysis for unreplicated fractional factorials. *Technometrics.* (1986);28:11–18.
11. Dean A. M., Lewis S. M. *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*. Springer, New York. (2006).
12. Overstall A. M., Woods D. C. Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model. *J. Roy. Statist. Soc. C*. (2016).
13. Bandura R. *Composite indicators and rankings: inventory* 2011. Tech. rep., United Nations Development Programme – Office of Development Studies.
14. Maozhun S., Ji L. Improved Garson algorithm based on neural network model. *Proceedings of the 29th Chinese Control and Decision Conference, CCDC*. (2017).
15. Блюмин С. Л., Суханов В. Ф., Чеботарев С. В. *Экономический факторный анализ*. Липецк: Издательство ЛЭГИ. (2004).
16. Blyumin S. L., Borovkova G. S., Serova K. V., Sysoev A. S. Analysis of finite fluctuations for solving big data management problems. *Application of Information and Communication Technologies (AICT), 9th International Conference*. (2015).
17. Сысоев А. С., Шеглеватых Р. В. Методы обнаружения аномалий в больших данных медицинской природы. *Современные сложные системы управления HTCS'2018: сборник трудов XIII Международной научно-практической конференции*. (2018).
18. Sysoev A., Ciurlia A., Sheglevatych R., Blyumin S. Sensitivity Analysis of Neural Network Models: Applying Methods of Analysis of Finite Fluctuations. *Periodica Polytechnica Electrical Engineering and Computer Science* [Internet]. (2019) [cited 2020];63(4):306-11. Available from: https://pp.bme.hu/eecs/article/view/14654

REFERENCES

1. Parmar J. D., Patel J. T. Anomaly Detection in Data Mining: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*. (2017);7(4):32-40.
2. Liu F. T., Ting K. M., Zhou Z. H. Isolation forest . *IEEE International Conference on Data Mining, ICDM* (2008).
3. Liu F. T., Ting K. M., Zhou Z. H. Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data.* (2012).
4. Rosenwasser E. N., Yusupov R. M. *Sensitivity of Control Systems*. Moscow: Nauka. (1981).
5. Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M., Tarantola S. *Global Sensitivity Analysis – The Primer*. John Wiley & Sons, Hoboken (2008).
6. Saltelli A., Tarantola S., Campolongo F. Sensitivity analysis as an ingredient of modeling. *Stat. Sci*. (2000); 15(4):377–395.
7. Kurowicka D., Cooke R. *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley. (2006).
8. Cacuci D. G. *Sensitivity and Uncertainty Analysis Theory*. Chapman and Hall/CRC. (2005).
9. Hamby D. M. A review of techniques for parameter sensitivity analysis of environmental models. *Environ. Monit. Assess*. (1994); 32(2):135–154.
10. Box G. E. P., Meyer R. D. An analysis for unreplicated fractional factorials. *Technometrics*. (1986);28:11–18.
11. Dean A. M., Lewis S. M. *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*. Springer, New York. (2006).
12. Overstall A. M., Woods D. C. Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model. *J. Roy. Statist. Soc. C*. (2016).
13. Bandura R. *Composite indicators and rankings: inventory* 2011. Tech. rep., United Nations Development Programme – Office of Development Studies.
14. Maozhun S., Ji L. Improved Garson algorithm based on neural network model. *Proceedings of the 29th Chinese Control and Decision Conference, CCDC*. (2017).
15. Blyumin S. L., Sukhanov V. F., Chebotarev S. V. *Economic Factor Analysis*. Lipetsk: LEGI Publishing. (2004).
16. Blyumin S. L., Borovkova G. S., Serova K. V., Sysoev A. S. Analysis of finite fluctuations for solving big data management problems. *Application of Information and Communication Technologies (AICT), 9th International Conference*. (2015).
17. Sysoev A. S., Scheglevatych R. V. Approaches to Detect Anomalies in Big Data of Medical Nature. *Proceedings of International Conference HTCS2018*. (2018).
18. Sysoev A., Ciurlia A., Sheglevatych R., Blyumin S. Sensitivity Analysis of Neural Network Models: Applying Methods of Analysis of Finite Fluctuations. *Periodica Polytechnica Electrical Engineering and Computer Science* [Internet]. (2019) [cited 2020]; 63(4):306-11. Available from: https://pp.bme.hu/eecs/article/view/14654

ИНОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Щеглеватых Роман Вячеславович, Федеральный фонд обязательного медицинского страхования Москва, Россия, Начальник Управления информационно-аналитических технологий.
e-mail: scheg l111@mail.ru

Roman V. Scheglevatych, Head of Data Processing Department, Federal Fund of Compulsory Medical Insurance, Moscow, Russian Federation

Сысоев Антон Сергеевич, канд. техн. наук, доцент, кафедра прикладной математики, ФГБОУ ВО "Липецкий государственный технический университет" Факультет автоматизации и информатики, Липецк, Российская Федерация.
e-mail: sysoev_as@stu.lipetsk.ru
ORCID: 0000-0002-0866-1124

Anton S. Sysoev, Cand. Sci. (Techn.), Associate Professor, Department of Applied Mathematics, Federal State Budgetary Educational Institution of Higher Education "Lipetsk State Technical University", Lipetsk, Russian Federation