DOI: 10.15593/2499-9873/2021.1.02
УДК 519.25: 004.891.3

Р.В. Щеглеватых, А.С. Сысоев

Липецкий государственный технический
университет, Липецк, Россия

ИССЛЕДОВАНИЕ НЕЙРОСЕТЕВОЙ МОДЕЛИ ОБНАРУЖЕНИЯ АНОМАЛЬНЫХ НАБЛЮДЕНИЙ В МАССИВАХ ДАННЫХ

Цифровизация различных сфер экономической и социальной деятельности сопровождается возникновением больших массивов данных, обрабатывая которые необходимо выявлять определенные зависимости, строить модели процессов и систем. Посвящена разработке и исследованию математической модели классификации данных о фактах оказания медицинской помощи в учреждениях Липецкой области. В качестве массива входных данных использованы индикаторы оказания медицинской помощи, разделенные на пять групп (данные, характеризующие пациента; данные, характеризующие медицинское учреждение, в котором была оказана помощь; индикаторы заболевания; данные о медицинском сотруднике, оказавшем помощь; индикаторы, характеризующие специфические особенности посещения пациентом конкретного специалиста). Объем записей, на которых проводилось исследование, – более одного миллиона записей о фактах оказания помощи населению. Цель исследования – предложить модели и подходы к выявлению ошибочных записей, а также случаев фальсификации.

Приводится постановка задачи бинарной классификации. 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 модели обнаружения аномальных значений фиксируемого показателя на основе комбинации алгоритма изолирующего леса для оценки показателя аномальности наблюдения и последующего применения нейросетевого классификатора.

Исследование содержит результаты вычислительных экспериментов по определению порогового значения для разделения записей на классы аномальных наблюдений и данные, не обладающие признаками аномальности.

Для оценки того, какие факторы должны быть переданы на вход нейросетевого классификатора (с целью повышения временной эффективности обработки данных), был синтезирован подход к редукции нейросетевой модели, основанный на анализе чувствительности. Классическим подходом при рассмотрении чувствительности систем является нахождение чувствительности по параметрам изучаемой системы, однако существует и направление анализа чувствительности, предполагающее использование в качестве оцениваемых параметров системы ее факторы. Предлагаемый подход к анализу чувствительности модели по факторам основан на использовании анализа конечных изменений. В основе такого анализа – замена математической модели зависимости выхода системы от факторов на модель зависимости конечного изменения выхода от конечных изменений факторов. Из математического анализа известна такая структура – это теорема Лагранжа о промежуточной точке. Подход позволяет определить значения так называемых факторных нагрузок. Приводится подход к усреднению полученных значений факторных нагрузок и построению интервальных характеристик для их оценивания. Приводится исследование устойчивости предлагаемой процедуры вычисления коэффициентов чувствительности модели.

Ключевые слова: математическая модель, анализ чувствительности, обнаружение аномалий, необычные наблюдения, выбросы, анализ конечных изменений, теорема Лагранжа о промежуточной точке, классификатор, бинарная классификация, нейросетевые модели.

R.V. Sheglevatych, A.S. Sysoev

Lipetsk State Technical University, Lipetsk, Russian Federation

STUDY ON NEURAL NETWORK MODEL TO DETECT ANOMALIES IN DATASETS

Digitalization of various spheres of economic and social life is accompanied by the emergence of large amounts of data, processing of which is necessary to identify certain dependencies, build models of processes and systems. The study is devoted to the development and research of a mathematical model for the classification of data on medical care in the medical organization of Lipetsk region. As inputs there were used indicators of medical care, divided into five groups (data describing patient; data describing the medical organization in which the care was provided; indicators of the disease; data on health employee that assisted; indicators characterizing the specific features of the patient's visits to a particular specialist). The volume of records on which the study was conducted is more than one million records of the facts. The purpose of the study is to propose models and approaches for identifying erroneous records, as well as cases of falsification.

The paper presents a statement of the binary classification problem. Anomaly detection refers to the problem of finding data that does not correspond to some expected process behavior or indicator that occurs in the system. When building systems for detecting anomalous observations, much attention must be paid to the model underlying the system. The study is devoted to the construction of a model for detecting anomalous values of a fixed indicator based on a combination of an isolation forest algorithm to estimatie the observation anomaly index and the subsequent application of a neural network classifier.

The study contains the results of computational experiments to determine the threshold value for dividing records into classes of anomalous observations and data that do not have signs of abnormality.

To evaluate which factors should be passed to the input of the neural network classifier (in order to increase the time efficiency of data processing), the approach to the reduction of the neural network model based on Sensitivity Analysis was proposed. The classical approach when considering the sensitivity of systems is to find the sensitivity by the parameter of the system under study, however, there is also a direction of Sensitivity Analysis that involves using its factors as the estimated parameters of the system. The proposed approach is based on applying Analysis of Finite Fluctuation. This analysis is based on replacing the mathematical model of the dependence of the system output on factors with a model of the dependence of the finite fluctuation in output on the finite fluctuations of factors. In Mathematical Analysis such a structure is known – this is Lagrange mean value theorem. The approach allows us to determine the values of the so-called factor loads. The paper presents a new approach to averaging the obtained values of factor loads and constructing interval characteristics for their estimation. A study of the stability of the proposed procedure for calculating the sensitivity coefficients of the model is presented.

Keywords: mathematical model, sensitivity analysis, anomaly detection, abnormal observations, outliers, finite fluctuations analysis, Lagrange mean value theorem, classifier, binary classification, neural network models.

Введение

Задачи математического моделирования технических, социальных, экономических систем или технологических процессов требуют высокой степени уверенности в достоверности и качестве входной информации, используемой как для их структурной, так и для параметрической идентификации. В ходе синтеза модели, ее анализа или практического использования важным условием является выявление наблюдений, которые не могут быть классифицированы как нормальные, т.е. которые не подчиняются законам системы или процесса, требуют де-

тального исследования, могут оказать пагубное влияние на результат, полученный с помощью модели. Одним из хорошо зарекомендовавших себя инструментов моделирования, а также численного анализа систем, являются искусственные нейронные сети. Однако для повышения точности моделирования и возможности частичной интерпретации результатов моделирования актуальной становится задача выбора наиболее влиятельных входов нейросетевой модели, используемая впоследствии и в алгоритме выявления аномальных наблюдений. Решение этой задачи лежит в области анализа чувствительности по факторам математической модели, одной из целей которого как раз и является редукция моделей. В случае исследования нейросетевых моделей применяют алгоритмы, предполагающие объяснение характеристик нейронной сети через анализ ее весовых коэффициентов. Однако в силу существования различных параметров нейронной сети, доставляющих схожие выходы, такие алгоритмы не являются устойчивыми. Для решения задачи анализа чувствительности предлагается использовать известный метод анализа конечных изменений, основанный на применении теоремы Лагранжа о промежуточной точке и рассматривающий некоторые конечные изменения факторов модели и их связь с изменением выхода модели. Особую значимость указанные подходы приобретают в решении задачи выявления аномальных наблюдений в зафиксированных данных об оказании медицинских услуг населению. Своевременное обнаружение таких записей позволяет вести оперативный контроль за качеством оказания медицинской помощи населению и способно минимизировать человеческие ошибки в данных, а также обнаруживать возможные фальсификации предоставленной информации.

1. Постановка задачи бинарной классификации

Рассматривается задача нахождения аномальных наблюдений в массивах данных. Пусть проведено наблюдение над n объектами, каждый из которых характеризуется m-мерным вектором признаков $X = (X_1, ..., X_m)$. Про каждый из объектов известно, что он принадлежит к одному из двух классов: K_1 – нормальных наблюдений или K_2 – аномальных наблюдений. Множество показателей X объектов классов K_1 и K_2 является обучающей выборкой.

Задача состоит в том, чтобы для каждого нового наблюдения $x = (x_1, ..., x_m)$ вектора признаков с учетом совершения некоторой

ошибки второго рода (возможность того, что доля нормальных наблюдений будет отнесена к аномальным) определить класс объекта K_i, $i = 1, 2$, к которому его следует отнести. Необходимо синтезировать решающее правило (с учетом ошибки первого рода) $h(x): X \to \{d_1, d_2\}$, где $h(s) = d_1 \Leftrightarrow s \in K_1$.

Для того чтобы отнести новое наблюдение к одному из возможных классов, определим дискриминантную пороговую функцию $t(x): X \to R$. Тогда классификатор примет вид

$$h^{\lambda}(x) = \begin{cases} d_1, & t(x) \geq \lambda, \\ d_2, & t(x) < \lambda. \end{cases}$$

В заданном решающем правиле в качестве функции $t(x)$ возможно использование различных подходов. Далее для решения поставленной задачи в качестве указанной функции используется выходное значение нейросетевой модели.

2. Определение порогового значения разделения наблюдений

В исследовании качество классификации (качество модели) оценивалось на основе точности и полноты, а также комбинации указанных характеристик [1]. Точность классификации $P(h)$ определяется как доля объектов, которые были распознаны как объекты класса аномальных наблюдений. Полнота классификации $R(h)$ показывает, какая доля объектов, реально относящаяся к классу аномальных наблюдений, была предсказана. Часто используют среднее гармоническое точности и полноты (F-меру) [2], которая определяется как

$$F(h) = \frac{2P(h)R(h)}{P(h) + R(h)}.$$

Одним из способов оценки порогового значения для бинарной классификации, построенного в соответствии с описанным выше алгоритмом, является использование кривых ошибок (или ROC-кривых) [3]. ROC-кривая представляет собой график зависимости полноты классификации $R(h)$ от единицы минус величина специфичности (отношение количества истинно отрицательных решений к сумме ис-

тинно отрицательных и ложноположительных решений) по всем возможным пороговым значениям λ.

3. Модели обнаружения аномальных наблюдений

Одной из групп подходов к обнаружению аномалий являются алгоритмы, основанные на применении методов классификации [4]. В качестве инструментов для классификации могут быть использованы различные структуры: деревья решения, модели нечеткой логики, наивные байесовские модели, генетические алгоритмы, нейронные сети, опорные векторы и т.п. 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 не на построении модели, определяющей «нормальное» значение и отвергающей все не 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 [5, 6]. 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 кон-

тролирующими специалистами. Для повышения качества обнаружения аномальных значений [7] 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 нетипичные по сравнению с ближайшими данными), а затем, применяя нейросетевой классификатор [8, 9], построить модель, способную находить выбросы для новых исходных данных:

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

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, основанный на применении анализа конечных изменений.

3.1. Анализ чувствительности

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 работе [10] как логичное расширение экономического факторного анализа и затем нашел применение в различных прикладных исследованиях [11]. 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

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

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

$$\mu(y) = \varphi(\mu(x_1),...,\mu(x_d)),$$

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 выполнить указанные преобразования. 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:

$$\Delta y = \sum_{i=1}^{d} \frac{\partial f(X^{(m)})}{\partial x_i} \cdot \Delta x_i,$$

$$X^{(m)} = \left(x_1^{(m)},...,x_d^{(m)}\right), \quad x_i^{(m)} = x_i^{(0)} + \alpha \Delta x_i,$$

$$i = 1,...,d, \quad 0 < \alpha < 1.$$

Промежуточная точка определяется значением параметра α.

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:

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

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

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

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.

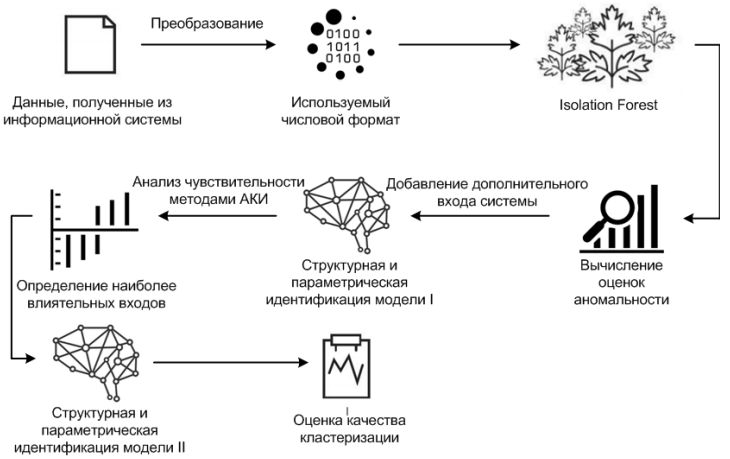

Рис. 1. Структурная схема подхода к обнаружению аномальных наблюдений в массивах данных

Структурная схема подхода к обнаружению аномальных наблюдений в массивах данных представлена на рис. 1. На начальном этапе полученные из информационной системы данные должны быть преобразованы в необходимый формат. Затем исходные данные подвергаются преобразованию алгоритмом изолирующего леса. Выбор именно этого алгоритма объясняется, во-первых, тем, что алгоритм распознает аномалии различных типов; во-вторых, сложность изолирующего дерева эффективнее большинства других алгоритмов; в-третьих, отсутствием па-

раметров выполнения. Вычисленные оценки аномальности для каждого из наблюдений формируют еще один вход системы, который затем будет использован при построении нейросетевой модели классификатора. Построенная модель, в свою очередь, подвергается описанному далее анализу чувствительности по входам для сокращения числа своих аргументов. По выбранным наиболее значимым входам происходит структурная и параметрическая идентификация новой нейросетевой модели классификатора с сокращенным числом входных переменных. На последнем этапе производится оценка качества классификации.

3.2. Точечные и интервальные оценки показателей чувствительности

В качестве устойчивой к выбросам в исходном наборе данных оценки среднего значения может быть использовано взвешенное среднее Тьюки. Алгоритм построения этой оценки носит итерационный характер и включает в себя следующие шаги [12]:

1. Вычисление среднего значения выборки (в начале работы алгоритма обычно используется медиана).

2. Определение расстояния от вычисленного среднего до каждого элемента выборки. В соответствии с этими расстояниями элементам выборки присваиваются различные веса, с учетом которых среднее значение пересчитывается. Характер весовой функции таков, что наблюдения, отстоящие от среднего достаточно далеко, не вносят большого вклада в значение взвешенного среднего.

Пусть A_{x_i} – выборка из рассчитанных факторных нагрузок для входа x_i, $A_{x_i} = \{A_{x_i}^1, ..., A_{x_i}^n\}$; $M_{A_{x_i}}$ – медиана выборки A_{x_i}; S – медиана выборки; $\left\{\left|A_{x_i}^1 - M_{A_{x_i}}\right|, ..., \left|A_{x_i}^n - M_{A_{x_i}}\right|\right\}$ – абсолютное отклонение среднего. Для каждого элемента $A_{x_i}^k$ ($k = 1,\ldots,n$) выборки A_{x_i} вычисляют отклонение от среднего

$$u_k = \frac{A_{x_i}^k - M_{A_{x_i}}}{cS + \xi},$$

где c – параметр, определяющий, насколько оценка чувствительна к выбросам; ξ – малая величина, основное назначение которой – исключить возможность деления на ноль.

Для нахождения веса каждого наблюдения выборки используется биквадратная функция вида

$$w(u) = \begin{cases} (1-u^2)^2, & |u| \le 1, \\ 0, & |u| > 1. \end{cases}$$

Оценка взвешенного Тьюки

$$T_{A_{x_i}} = \frac{\sum_{k=1}^{n} w(u_k) A_{x_i}}{\sum_{k=1}^{n} w(u_k) \cdot}.$$

Помимо точечной оценки среднего находят интервал для построения величины при помощи приближения распределением Стьюдента.

Симметричный $(1-\alpha)$%-ный доверительный интервал задается формулой

$$T_{A_{x_i}} \pm t_{df}^{(1-\alpha/2)} \cdot \frac{S_{A_{x_i}}}{\sqrt{n}},$$

$$S_{A_{x_i}} = \sqrt{n} \cdot \frac{\sqrt{\sum_{|u_k| \le 1} (A_{x_i}^k - T_{x_i})^2 (1-u_k^2)^4}}{\left| \sum_{|u_k| \le 1} (1-u_k^2)(1-5u_k^2) \right|},$$

где $t_{df}^{(1-\alpha/2)}$ – $(1-\alpha/2)$-квантиль распределения Стьюдента с числом степеней свободы $df = \max(0, 7(n-1), 1)$.

3.3. Устойчивость метода анализа чувствительности

Входы системы могут иметь неустранимые погрешности, которые тем не менее не должны вызывать неточностей при вычислениях оценок влияния факторов на выход системы. Устойчивость численного метода характеризуется небольшими отклонениями выходного значения при незначительных изменениях во входах. Пусть в результате решения задачи по входному значению фактора x_i находится значение величины выхода. Входная величина x_i имеет некоторую погрешность εx_i, выход y имеет погрешность εy, значит, если при $\varepsilon x_i \to 0$, $i = 1,\ldots,n$, $\varepsilon y \to 0$, метод является вычислительно устойчивым.

Для практического исследования устойчивости необходимо провести серию вычислительных экспериментов; полученные соответствующие выборки, состоящие из оценок значимости входов, исследовать с точки зрения схожести в смысле средних значений. Для этого можно сформулировать нулевую гипотезу о сдвиге средних значений выборок при альтернативной гипотезе об отсутствии такого сдвига, проверка которой сводится в дальнейшем к вычислению некоторой статистики (например, непараметрической статистики Манна – Уитни – Вилкоксона) [13].

Пусть $x_{11},...,x_{1n}$ и $x_{21},...,x_{2m}$ – упорядоченные по возрастанию выборки. Для проверки гипотезы вычисляют

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}, \quad \text{где} \quad h_{ij} = \begin{cases} 1, & x_{1i} < x_{2j}, \\ 0, & x_{1i} > x_{2j}. \end{cases}$$

U-статистика Манна – Уитни определяет точное число пар значений x_{1i} и x_{2j}, для которых $x_{1i} < x_{2j}$. С *U*-статистикой связана статистика Вилкоксона, определяемая суммой рангов элементов одной выборки (например, x_1 объемом *n*) в общей упорядоченной последовательности элементов совместной выборки (объемом *m* + *n*):

$$R = mn + \frac{n(n+1)}{2} - U.$$

При *m*, *n* > 20 применима аппроксимация

$$W = \frac{R - \dfrac{n(n+m+1)}{2}}{\sqrt{\dfrac{nm(n+m+1)}{12}}}.$$

Статистика *W* аппроксимируется нормальным распределением, гипотеза сдвига (несовпадения средних выборок) отклоняется с достоверностью α, если $|W| > u_{\frac{1+\alpha}{2}}$.

4. Вычислительные эксперименты. Полученные результаты

Прикладная задача заключается в нахождении аномальных значений среди данных, полученных из Единой государственной инфор-

мационной системы в сфере здравоохранения. Стоит отметить, что операции над данными выполнялись после их обезличивания. Под найденными аномальными значениями понимаются записи, хранящиеся в системе, по своей форме, структуре или содержанию заметно отличающиеся от нормального формата или содержания. Появление и существование таких записей может быть связано, во-первых, с человеческим фактором (ввиду большого количества данных по каждому факту оказания медицинской помощи персонал может совершать технические ошибки) и, во-вторых, с возможными подлогами со стороны медицинского персонала (в ситуации страховой медицины не исключены умышленные искажения медицинских записей с целью получения выгоды). В качестве исходных данных в исследовании выступал набор данных по оказании помощи населению Липецкой области, который включает в себя множество показателей, разделенных на пять групп: данные, характеризующие пациента; данные, характеризующие медицинское учреждение, в котором была оказана помощь; индикаторы заболевания; данные о медицинском сотруднике, оказавшем помощь; индикаторы, характеризующие специфические особенности посещения пациентом конкретного специалиста. Данные были собраны с февраля по май 2019 г. и содержат более одного миллиона случаев. Следует отметить, что один пациент мог быть связан со многими записями (по его визитам к врачу). Не все индикаторы использованы для обучения нейросетевого классификатора, какие-то – в силу своего значения (например, наименование медицинской организации, к которой прикреплен пациент, или наименование структурного подразделения медицинской организации, в котором пациент получал медицинскую помощь), другие – в силу невозможности их использования в исходном формате (например, не столько важны даты начала и окончания лечения, сколько продолжительность оказанных процедур). Таким образом, для обучения нейросетевого классификатора на начальном этапе было использовано 27 входов [14, 15].

Была использована нейросетевая модель, имеющая структуру

$$y_k = \psi_1\left(b_0 + \sum_{j=1}^{27} w_j \psi_2\left(b_1 + \sum_{m=1}^{27} w_{jm} x_m\right)\right),$$

где y_k – модельные значения; $x_m \in \mathbf{x}$ – факторы системы; w_j и w_{jm} – ве-

совые коэффициенты; b_0 и b_1 – свободные коэффициенты нейронов выходного и скрытого слоев соответственно; $\psi_1(\text{net}) = \psi_2(\text{net}) = 1/(1+\exp(-\text{net}))$ – логистические функции активации.

На рис. 2 представлены полученные по предложенному подходу оценки коэффициентов чувствительности модели с учетом вычисленных границ. Стоит отметить, что представлена информация о коэффициентах чувствительности без учета направления их влияния (использованы абсолютные значения найденных показателей).

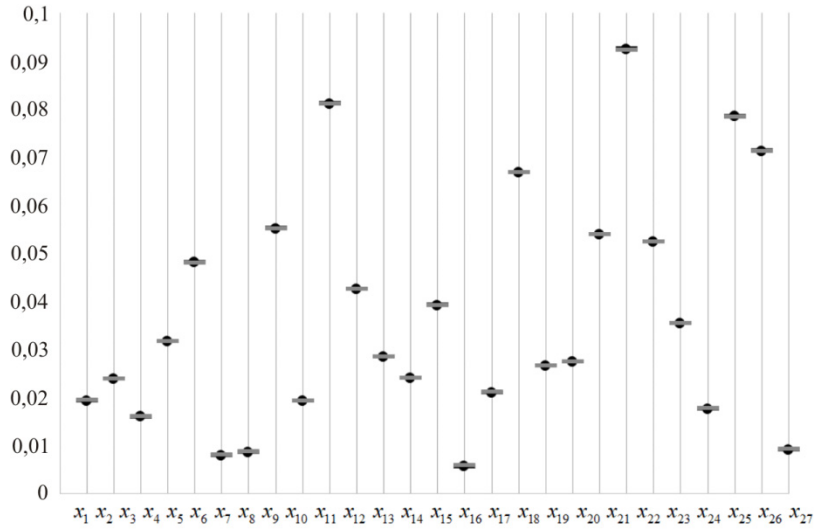

Рис. 2. Результаты анализа чувствительности модели

Для выбора рационального значения порога с целью последующего его применения для классификации новых реализаций были проведены серии вычислительных экспериментов. Среднемесячный объем записей о фактах оказания медицинской помощи составляет около полумиллиона записей. Обучение классификатора с нейросетевой структурой с учетом проведенного анализа чувствительности для сокращения входов модели проходило на выборке объемом 553 738 реализаций. Определение порогового значения по ROC-кривым проходило на выборках объемом 10, 25, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500 тысяч реализаций. Было проведено исследование зависимости доли правильно классифицированных значений (рис. 3) и *F*-меры (рис. 4) от объема выборки. Установлено, что наиболее рациональным являет-

ся выбор значения λ = 0,01 494 603 при нормированных значениях входных переменных модели.

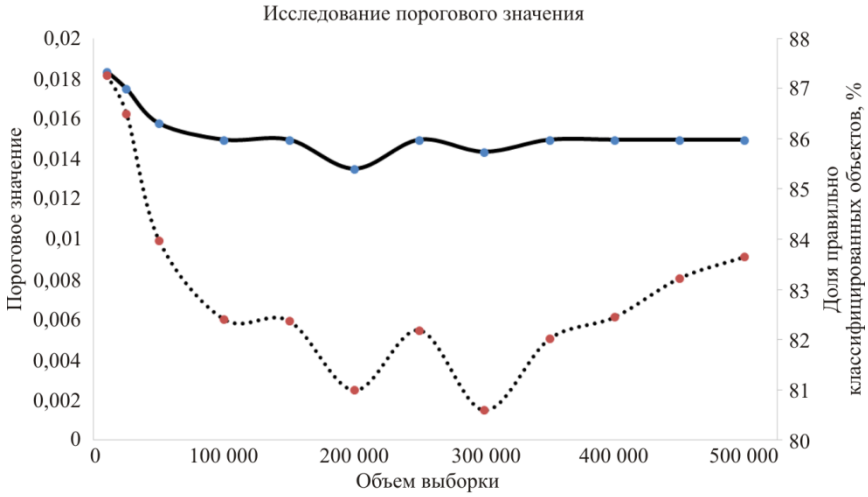

Рис. 3. Зависимость доли правильно классифицированных значений (·●·) и порогового значения (–●–) от объема выборки

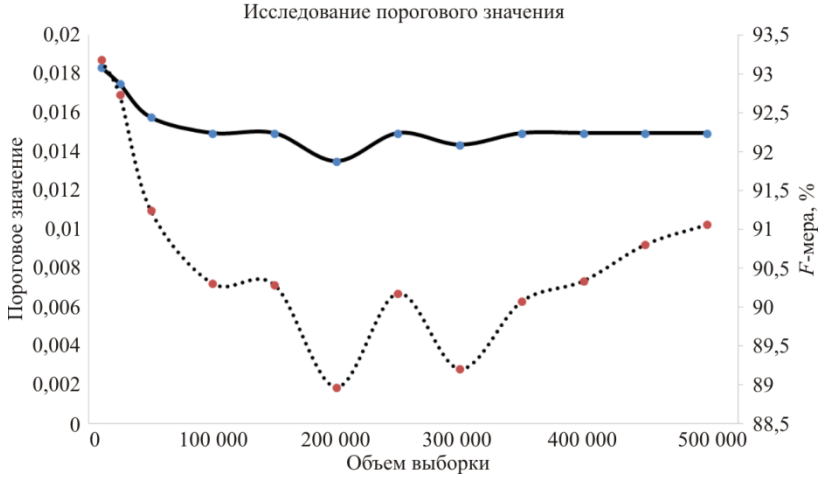

Рис. 4. Зависимость *F*-меры (·●·) и порогового значения (–●–) от объема выборки

Заключение

В работе представлен подход к построению бинарного классификатора для нахождения аномальных записей в массивах данных об оказании медицинских услуг населению. Особое внимание уделено анализу

чувствительности нейросетевых моделей, который основан на анализе конечных изменений. Следует отметить, что этот подход к анализу чувствительности не ограничивается только одним классом нейросетевых моделей. Рассматриваемая модель должна быть дифференцируемой функцией для применения ключевой теоремы подхода (теоремы Лагранжа о среднем). В отличие от известных подходов (например, использование коэффициентов чувствительности Соболя), предложенный подход не основан на аппроксимации статистических параметров исследуемой структуры и, в отличие от стратегии Гарсона (подход, используемый для оценки чувствительности входов нейронных сетей), оперирует как параметрами, так и факторами исследуемой модели.

Список литературы

1. Bramer M. Estimating the predictive accuracy of a classifier // Principles of Data Mining. – 4th ed. – London: Springer-Verlag London, 2020. – P. 79–92.

2. Sawade C., Landwehr N., Scheffer T. Active estimation of f-measures // Advances in Neural Information Processing Systems. – 2010. – Vol. 2. – P. 2083–2091.

3. Obuchowski N.A. Nonparametric analysis of clustered ROC curve data // Biometrics. – 1997. – Vol. 53, no. 2. – P. 567–578.

4. 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. – Vol. 7, iss. 4. – P. 32–40.

5. Liu F.T., Ting K.M., Zhou Z.H. Isolation forest // 2008 Eight IEEE International Conference on Data Mining (ICDM), Pisa, Italy. 15–19 December 2008. – Los Alamitos: IEEE, 2008. – Art. 10472172. – P. 413–422. DOI: 10.1109/ICDM.2008.17

6. Liu F.T., Ting K.M., Zhou Z.H. Isolation-based anomaly detection // ACM Transactions on Knowledge Discovery from Data. – 2012. – Vol. 6, iss. 1. – Art. 3. – P. 1–39. DOI: 10.1145/2133360.2133363

7. Щеглеватых Р.В., Сысоев А.С. Математическая модель обнаружения аномальных наблюдений с использованием анализа чувствительности нейронной сети [Электронный ресурс] // Моделирование, оптимизация и информационные технологии. – 2020. – Т. 8, № 1. – 14 с. – URL: https://moit.vivt.ru/wp-content/uploads/2020/02/ScheglevatychSysoev_1_20_1.pdf DOI: 0.26102/2310-6018/2020.28

8. Sysoev A.S., Blyumin S.L., Scheglevatych R.V. Approach to sensitivity analysis of neural network models based on analysis of finite fluctuations // 14th International Conference on Pattern Recognition and Information Processing (PRIP'2019), Minsk, Belarus. 21–23 May 2019 / Belarus State University of Informatics and Radioelectronics. – Minsk, 2019. – P. 97–100.

9. Sysoev A., Scheglevatych R. Combined approach to detect anomalies in health care datasets // 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russia, 20–22 November 2019. – Los Alamitos: IEEE, 2019. – P. 359–363. DOI: 10.1109/SUMMA48161.2019.8947605

10. Блюмин С.Л., Суханов В.Ф., Чеботарев С.В. Экономический факторный анализ: монография / ЛЭГИ. – Липецк, 2004. – 148 с.

11. Analysis of finite fluctuations for solving big data management problems / S.L. Blyumin, G.S. Borovkova, K.V. Serova, A.S. Sysoev // 2015 9th International Conference on Application of Information and Communication Technologies (AICT), Rostov on Don, Russia, 14–16 October 2015. – Los Alamitos: IEEE, 2015. – Art. 15620282. – 4 p. DOI: 10.1109/ICAICT.2015.7338514

12. Hoaglin D.C., Mosteller F., Tukey J.W. Understanding robust and exploratory data analysis. – New York: Wiley-Interscience, 2000. – 472 p.

13. Mann H.B., Whitney D.R. On a test of whether one of two random variables is stochastically larger than the other // Annals of Mathematical Statistics. – 1947. – Vol. 18, no. 1. – P. 50–60.

14. Sensitivity analysis of neural network models: applying methods of analysis of finite fluctuations / A. Sysoev, A. Ciurlia, R. Sheglevatych, S. Blyumin // Periodica Polytechnica Electrical Engineering and Computer Science. – 2019. – Vol. 63, iss. 4. – P. 306–311. DOI: 10.3311/PPee.14654

15. Sheglevatych R.V., Sysoev A.S. Analysis of finite fluctuations as a basis of defining a set of neural network model inputs // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems: сб. науч. тр. / под ред. В.В. Голенкова (гл. ред.) [и др.]; Белорус. гос. ун-т информ. и радиоэлектрон. – Минск, 2020. – Вып. 4. – С. 313–316.

References

1. Bramer M. Estimating the Predictive Accuracy of a Classifier. Principles of Data Mining, Springer-Verlag London, 2020, pp. 79–92.

2. Sawade C., Landwehr N., Scheffer T. Active estimation of f-measures. *Advances in Neural Information Processing Systems*, 2010, vol. 2, pp. 2083–2091.

3. Obuchowski N. A. Nonparametric analysis of clustered ROC curve data. *Biometrics*, 1997, vol. 53, no. 2, pp. 567–578.

4. 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, vol. 7, iss. 4, pp. 32–40.

5. Liu F.T., Ting K.M., Zhou Z.H. Isolation forest. *IEEE International Conference on Data Mining (ICDM)*, IEEE, 2008, art. 10472172, pp. 413–422, DOI: 10.1109/ICDM.2008.17

6. Liu F.T., Ting K.M., Zhou Z.H. Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, 2012, vol. 6, iss. 1, art. no. 3, pp. 1–39, DOI: 10.1145/2133360.2133363

7. Shcheglevatykh R.V., Sysoev A.S. Matematicheskaia model' obnaruzheniia anomal'nykh nabliudenii s ispol'zovaniem analiza chuvstvitel'nosti neironnoi seti [Mathematical model to detect anomalies using Sensitivity Analysis applying to neural network]. *Modeling, Optimization and Information Technology*, 2020, vol. 8, iss. 1, available at: https://moit.vivt.ru/wp-content/uploads/2020/02/ScheglevatychSysoev_1_20_1.pdf. DOI: 0.26102/2310-6018/2020.28

8. Sysoev A.S., Blyumin S.L., Scheglevatych R.V. Approach to Sensitivity Analysis of Neural Network Models Based on Analysis of Finite Fluctuations. 14th International Conference on Pattern Recognition and Information Processing (PRIP'2019), Minsk, Belarus state university of informatics and radioelectronics, 2019, pp. 97–100.

9. Sysoev A., Scheglevatych R. Combined Approach to Detect Anomalies in Health Care Datasets. 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA), IEEE, 2019, pp. 359–363. DOI: 10.1109/SUMMA48161.2019.8947605

10. . Blyumin S.L., Sukhanov V.F., Chebotarev S.V. Ekonomicheskii faktornyi analiz [Economic factor analysis]. Lipetsk, Lipetsk ecology institute, 2004, 148 c

11. Blyumin S.L., Borovkova G.S., Serova K.V., Sysoev A.S. Analysis of finite fluctuations for solving big data management problems. 2015 9th International Conference on Application of Information and Communication Technologies (AICT), IEEE, 2015. – 4 p. – art. 15620282. DOI: 10.1109/ICAICT.2015.7338514

12. Hoaglin D. C., Mosteller F., Tukey J. W. Understanding robust and exploratory data analysis, NewYork, Wiley-Interscience, 2000, 472 p.

13. Mann H.B., Whitney D.R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 1947, vol. 18, no. 1, pp. 50–60.

14. 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*, 2019, vol. 63, iss. 4, pp. 306–311. DOI: 10.3311/PPee.14654

15. Sheglevatych R.V., Sysoev A.S. Analysis of Finite Fluctuations as a Basis of Defining a Set of Neural Network Model Inputs. Open Semantic Technologies for Intelligent Systems, Minsk, Belarus state university of informatics and radioelectronics 2020, Iss. 4, pp. 313–316.

Статья получена: 28.01.2021

Статья принята: 02.03.2021

Сведения об авторах

Щеглеватых Роман Вячеславович (Липецк, Россия) – соискатель кафедры «Прикладная математика», Липецкий государственный технический университет (398055, Липецк, ул. Московская, 30, e-mail: schegl111@mail.ru).

Сысоев Антон Сергеевич (Липецк, Россия) – кандидат технических наук, доцент, доцент кафедры «Прикладная математика», Липецкий государственный технический университет (398055, Липецк, ул. Московская, 30, e-mail: sysoev_as@stu.lipetsk.ru).

About the authors

Roman V. Scheglevatych (Lipetsk, Russian Federation) – Postgraduate Student, Department of Applied Mathematics, Lipetsk State Technical University (30, Moskovskaya st., Lipetsk, 398055, e-mail: schegl111@mail.ru).

Anton S. Sysoev (Lipetsk, Russian Federation) – Ph.D. in Engineering, Associate Professor, Department of Applied Mathematics, Lipetsk State Technical University (30, Moskovskaya st., Lipetsk, 398055, e-mail: sysoev_as@stu.lipetsk.ru).

Библиографическое описание статьи согласно ГОСТ Р 7.0.100–2018:

Щеглеватых, Р.В. Исследование нейросетевой модели обнаружения аномальных наблюдений в массивах данных / Р. В. Щеглеватых, А. С. Сысоев. – текст : непосредственный – DOI 10.15593/2499-9873/2021.1.02 // Прикладная математика и вопросы управления = Applied Mathematics and Control Sciences. – 2021. – № 1. – С. 23–40.

Цитирование статьи в изданиях РИНЦ:

Щеглеватых Р.В., Сысоев А.С. Исследование нейросетевой модели обнаружения аномальных наблюдений в массивах данных // Прикладная математика и вопросы управления. – 2021. – № 1. – С. 23–40. – DOI: 10.15593/2499-9873/2021.1.02

Цитирование статьи в references и международных изданиях:

Cite this article as:

Sheglevatych R.V., Sysoev A.S. Study on neural network model to detect anomalies in datasets. *Applied Mathematics and Control Sciences*, 2021, no. 1, pp. 23–40. DOI: 10.15593/2499-9873/2021.1.02 (*in Russian*)