# Analysis of Finite Fluctuations as a Basis of Defining a Set of Neural Network Model Inputs

Roman Sheglevatych
*Lipetsk Region Health Department*
Lipetsk, Russia
schegl111@mail.ru

Anton Sysoev
*Department of Applied Mathematics*
*Lipetsk State Technical University*
Lipetsk, Russia
sysoev_as@stu.lipetsk.ru

***Abstract*—The paper describes an approach on the defining a set of neural network model inputs analyzing their influence on the output. The mathematical basis of such procedure is Analysis of Finite Fluctuations based on applying Lagrange mean value theorem. The applied problem under consideration in finding outliers in data from healthcare digital system records.**

***Keywords*—Analysis of Finite Fluctuations, neural networks, Sensitivity Analysis**

## I. Introduction. Outliers Detection Problem in Medical Healthcare Datasets

The very important problem which facing Russian healthcare system and consequently insurance companies is the qualitative way of organizing and functioning of healthcare single digital system [1]. It was announced that in two years every Russian will have a digital copy of his medical care story. In accordance, the approaches based on artificial intelligence will be widely implemented to help physicians on making decisions.

A part of the described challenge which is a peace of interest firstly of insurance companies and is a way to increase the quality of medical healthcare in general is the outliers detection in datasets obtained from medical healthcare digital systems.

Outliers (or anomalies) detection refers to the problem of finding data are not corresponding to some expected behavior of the process or indicator of the system [2]. When anomalies are detected, it is sometimes difficult to determine the threshold of the normality. Values which are close to threshold can be both normal and anomaly. That is why it is very important to verify each approach to outliers detection via expert estimates.

In the field of anomalies detection exist many fundamental and applied studies containing different approaches to solve the basic problem. They may be classified according the scheme used in the approach as supervised (cf. [3]), unsupervised and semi-supervised anomalies detection (cf. [4]). And also there is a classification which is based on the mathematical approach underlying used scheme.

This study is partially supported by Russian Foundation for Basic Research (RFBR) and Lipetsk regional administration. Grand No 19-47-480003 r_a.

## II. Problem to Define the Most Influenced Factors in Neural Network Model

Constructing a mathematical model of a technical, economic, social system or technological process, the question of choosing the most significant inputs affecting the response (output) of the studied structure is a relevant problem in information processing. Sensitivity Analysis allows to estimate the influence of the model output on its inputs, as well as to assess the importance of each factor. Being a universal mechanism to describe different complicated systems and processes, artificial neural networks are widely used in many areas of theoretical fields and practical applications.

Let $X = (x_1, ..., x_n)$ be system inputs (factors, variables), $G(\cdots)$ is an operator of the system and $Y$ is the system output (indicator), then the system can be represented as $Y = G(X)$.

In this applied study neural network was chosen as a structure of classifier. The most prominent approach to assess the sensitivity of neural network models is the Garson algorithm. This method is based on the study of weights constructed neural network model. It is believed that the variation of the studied coefficients can explain the characteristics of the "black box" neural network. According to the study [5], for three-layer neural network with a classical structure, factor sensitivity coefficients can be found as

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

where $i$, $j$ and $k$ are indexes for weights of input, hidden, and output layer weights respectively.

A well known approach which could be used in different cases (and does not depend on model type) is applying techniques in composite indicators [6], [7]. The idea behind these family of methods is to construct a composite indicator aggregating several factors with some weights, where weights define the degree of importance for each indicator. The most prominent approach in this family is Sobol sensitivity coefficients.

Table I
INDICATORS ON THE MEDICAL CARE RECORDS

| Type of the indicator | Code of the indicator | Explanation |
|---|---|---|
| Indicators belonging to the patient | CEL_OBSL | The purpose of the patient's appeal to the medical organization |
| | RSLT | The result of the medical care |
| | ISHOD | The patient's vitals |
| | voz | The patient's age |
| Indicators belonging to the medical organization providing medical care | lpu_p | The name of the medical organization to which the patient is assigned |
| | LPU | The name of the medical organization where the patient was treated |
| | KOD_TP | The code of the medical organization department where the patient was treated |
| | PODR | The name of the medical organization department where the patient was treated |
| | PROFIL | The profile of the provided medical care |
| | NAZ_PMP | The profile of the medical care to which the referral was given based on the results of the medical examination for patients of the 3rd group of health |
| | NAZ_PK | The profile of a round-the-clock or daily hospital place for which a referral for the hospitalization was given based on the results of medical examination for patients of the 3rd health group |
| | PROFIL_K | The profile of the place in a round-the-clock or day hospital where medical care was provided |
| Indicators belonging to the disease | IDCASE | The case unique identifier |
| | DS0 | The primary diagnosis |
| | DS1 | The basic diagnosis |
| | DS2 | The concomitant disease |
| | POVTOR | The sign of a treatment repeated case for a single disease |
| | TYPE_MN | The nature of the basic disease |
| | VIDTR | The sort of an injury |
| | KOD_KSG | The code of the clinical and statistical group of the disease in the conditions of a daily or round-the-clock hospital |
| Indicators belonging to the doctor | PRVS | The doctor's specialization |
| | SPEC_END | The regional localization of the doctor's specialization |
| | NAZ_SP | The specialization of the doctor to whom the appointment was made based on the results of medical examination for patients of the 3rd group of health |
| Indicators belonging to the particular case of the medical care | USL_OK | Conditions for the provided medical care |
| | VIDPOM | The type of medical care |
| | FOR_POM | The form how the medical care was provided |
| | DATE_1 | Start date of treatment * |
| | DATE_2 | End date of treatment * |
| | POL_VIS | The number of visits to a medical organization in a case of the medical care |
| | HOM_VIS | The number of home visits by the doctor to the patient in the case of the medical care |
| | ITAP | The stage of the medical examination or the preventive examination |
| | DISP | The type of the medical examination or the preventive examination |
| | RSLT_D | The result of the medical examination or the preventive examination |
| | OBR | The indicator characterizing the method of payment for medical care in case of outpatient treatment |
| | TIMEV | The time of the call to an ambulance * |
| | TIMEP | The time of the arrival of an emergency medical services * |
| | P_PER | The sign of the transfer to the daily department or to the round-the-clock department |
| | NAZR | The appointment of the doctor based on the results of medical examination for patients of the 3rd group of health |
| | NAZ_V | The type of examination assigned based on the results of the medical examination for patients of the 3rd group of health |
| | DN | The type of the dispensary observation |

Indicators marked with * are not analyzed directly, but used in some combinations.

### A. Analysis of Finite Fluctuations

Analysis of Finite Fluctuations (AFF) could be defined as an approach to analyze complicated system with the goal to build a dependency connecting finite fluctuations of a function and finite fluctuations of its arguments [8].

In a practical applications normally the argument $x$ could be measured, and its measurement $\mu(x)$ could have different forms, but the most used one to define the transition from initial value $x^{(0)} = a$ to the value $x^{(1)} = b$ is the absolute increment $\mu(x) = \Delta x = b - a$.

The main problem of AFF is formulated as follows. Let us have a model

$$y = f(X) = f(x_1, ..., x_n) \tag{1}$$

describing the connection between the response (model output) $y$ and its arguments (factors) $x_i$, $i = 1, ..., n$. It is necessary to transform Model (1) into the form

$$\mu y = \phi(\mu(x_1), ..., \mu(x_n)), \tag{2}$$

which shows the connection between the fluctuation of the output $\mu(y)$ and the fluctuations $\mu(x_i)$ of its factors $x_i$, $i = 1, ..., n$.

Mathematical Analysis gives in case of finite increments the model, which allows to represent (1) to an

exact connection between the finite fluctuation of the model output and its factors fluctuations (2). This is Lagrange mean value theorem (the formula of finite increments, intermediate value theorem of Differential Calculus) for multivariable functions, defined and continuous in a closed domain and having continuous partial derivatives inside this domain:

$$\Delta y = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\left(X^{(m)}\right) \cdot \Delta x_i, \tag{3}$$

where $X^{(m)} = \left(..., x_i^{(m)} = x_i^{(0)} + \alpha \cdot \Delta x_i, ...\right)$, $0 < \alpha < 1$. Here the mean (or intermediate) values of factors $x_i^{(m)}$ are defined by the value of parameter $\alpha$.

### B. Sensitivity Analysis based on AFF

Let us have a neural network model with $k$ hidden layers, which describes studied technical, economic or social system or technological process

$$Y^{(k)} = \Phi^{(k)}\Phi^{(k-1)}...\Phi^{(1)}X,$$

where $X = (x_1, ..., x_n)^T$.

In a current moment of time the initial vector of inputs is $X^{(0)} = \left(x_1^{(0)}, ..., x_n^{(0)}\right)^T$ and the value of system indicator is

$$Y_0^{(k)} = \Phi^{(k)}\Phi^{(k-1)}...\Phi^{(1)}\left(x_1^{(0)}, ..., x_n^{(0)}\right)^T.$$

After a while the inputs vector changes its value to $X^{(1)} = \left(x_1^{(1)}, ..., x_n^{(1)}\right)^T$ and system indicator is respectively

$$Y_1^{(k)} = \Phi^{(k)}\Phi^{(k-1)}...\Phi^{(1)}\left(x_1^{(1)}, ..., x_n^{(1)}\right)^T.$$

Thus, the difference (finite fluctuation) of the system indicator is

$$\Delta Y^{(k)} = Y_1^{(k)} - Y_0^{(k)} \tag{4}$$

and according to Lagrange mean value theorem (3) the same difference could be estimated as

$$\Delta Y^{(k)} = \sum_{i=1}^{n} \frac{\partial Y^{(k)}}{\partial x_i}\left(..., x_i^{(0)} + \alpha \cdot \Delta x_i\right) \cdot \Delta x_i. \tag{5}$$

Equating (4) and (5) and solving the resulting equation, it is possible to find the value $\alpha$ and then to estimate so called factor loads $A_{x_i}$:

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

$$= A_{x_1}\Delta x_1 + ... + A_{x_n}\Delta x_n.$$

For the approach based on AFF there are obtained $N-1$ estimates (because of existing $N-1$ finite fluctuations for $N$ input values data set) and their median values are taken as a sensitivity measure.

The described approach was tested on 2000 cases from the data set "neuraldat" from NeuralNetTools R package comparing with the other classically used approaches (Sobol sensitivity coefficients and Garson algorithm). All compared strategies gave similar results with a slight variation, which proves, that Analysis of Finite Fluctuations is not contradictory and could be applied in such kind of problems. But proposed approach has an undeniable advantage [9]. In contrast to Sobol sensitivity coefficients it does not use an approximation procedure to model statistical parameters of the studied structure and in contrast to Garson strategy, it operates with both parameters and factors of the studied model.

## III. Numerical Experiments

### A. Scope of Experiment. Data set Description

The data set on medical healthcare in Lipetsk region includes many indicators defined in Table I. Data were collected from February to May 2019 and show how medical care was provided in more than 1 billion cases. It should be noted, that one patient could be linked with many records (according to his visits to the physician). All the indicators could be divided into consolidated groups presented in Table I.

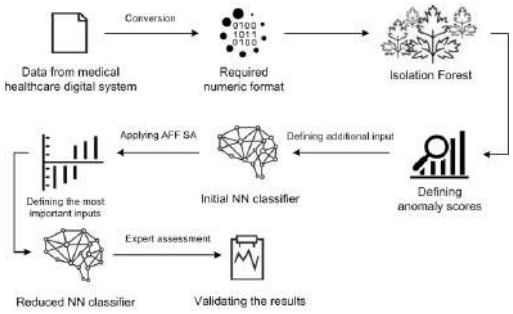


Figure 1. Work flow of the proposed solution

### B. Neural Network Classifier to Predict Outliers

As the neural network model has to be reduced according to its inputs, to improve the accuracy of the model it is proposed to add the results (anomaly score) of Isolation Forest algorithm applied to the inputs as an additional input. This step has positively established itself and is a basis of combined approach to detect anomalies in health care data sets described in [10]. The work flow of the proposed approach is given on Figure 1.

The initial neural network model with 34 input factors, 1 hidden layer with 3 neurons, and 1 output was investigated; the logistic activation functions were used, when the reduced one has 15 the most important inputs presented in Table III-A.

The data set used in the numerical experiment was divided into two subsets with 80% in training set and 20% in testing set. The final reduced model has demonstrated the general accuracy of 78% with 24% type I error of classification (false positive) and 16% type II error of classification.

Table II
INDICATORS WITH THE HIGHEST IMPACT ON THE NEURAL NETWORK OUTPUT

| Name of the indicator | Explanation | Degree of influence, % |
|---|---|---|
| lPU_P | The name of the medical organization to which the patient is assigned | 4,11 |
| USL_OK | Conditions for the provided medical care | 8,25 |
| SROK_LECH | The length of the treatment or of the hospitalization | 6,29 |
| CEL_OBSL | The purpose of the patient's appeal to the medical organization | 3,92 |
| PRVS | The doctor's specialization | 4,08 |
| SPEC_END | The regional localization of the doctor's specialization | 4,82 |
| POVTOR | The sign of a treatment repeated case for a single disease | 5,71 |
| TYPE_MN | The nature of the basic disease | 4,55 |
| ITAP | The stage of the medical examination or the preventive examination | 4,24 |
| RSLT_D | The result of the medical examination or the preventive examination | 4,77 |
| OBR | The indicator characterizing the method of payment for medical care in case of outpatient treatment | 3,15 |
| RAZN_SKOR | The time between calling to the ambulance and the arrival of medical service | 5,75 |
| VIDTR | The sort of an injury | 4,39 |
| NAZ_PK | The profile of a round-the-clock or daily hospital place for which a referral for the hospitalization was given based on the results of medical examination for patients of the 3rd health group | 5,33 |
| ANOMALY_SCORE | The anomaly score obtained by Isolation Forest algorithm applying | 2,85 |

## IV. RESULTS AND DISCUSSION

Analyzing results of the identified by the neural network model outliers it was found that in many cases detected results have deviations from ordinary records. Many of detected outliers could be assigned to one of the following groups. Firstly, in many founded records there were technical errors during filling the record like detecting that « dental consultation» was marked as «chronic disease». Secondly, in several cases there were too many visits to doctor with the minor ailments like visiting dentist 5 times to treat an enamel degradation.


## ACKNOWLEDGMENT

Authors are grateful to Lipetsk territorial compulsory medical insurance fund for the applied problem statement and access to the non-personal health care digital system records.


## REFERENCES


[1] The Federal law on 29.07.2017 no 242 "On amendments to certain legislative acts of the Russian Federation on the application of information technologies in the field of health protection."

[2] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," in Procedia Computer Science, 2015.

[3] Abe Naoki, Bianca Zadrozny, and John Langford. "Outlier detection by active learning," Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006.

[4] Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. "Distance-based outliers: algorithms and applications," The VLDB Journal 8.3-4, 2000.

[5] Maozhun Sun, and Liu Ji. "Improved Garson algorithm based on neural network model," 2017 29th Chinese Control And Decision Conference (CCDC). IEEE, 2017.

[6] Bandura Romina. "Composite indicators and rankings: Inventory 2011," New York: United Nations Development Programme, Office of Development Studies (UNDP/ODS Working Paper), 2011.

[7] A. Saltelli, S. Tarantola, and F. Campolongo. '"Sensitivity anaysis as an ingredient of modeling," Statistical Science 15.4, 2000.

[8] S.L. Blyumin, G.S. Borovkova, K.V. Serova and A.S. Sysoev "Analysis of Finite Fluctuations for Solving Big Data Management Problems," 9th Conference on Application of Information and Communication Technologies, (AICT). Rostov-on-Don, 2015, pp. 48-51.

[9] A. Sysoev, A. Ciurlia, R. Sheglevatych, and S. Blyumin, "Sensitivity Analysis of Neural Network Models: Applying Methods of Analysis of Finite Fluctuations", Periodica Polytechnica Electrical Engineering and Computer Science, vol. 63, no. 4, pp. 306-311, 2019.

[10] A. Sysoev, and R. Scheglevatych. "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.


## Анализ конечных изменений как основа формирования входов нейросетевой модели


Щеглеватых Р.В., Сысоев А.С.



В статье рассматривается подход к формированию набора входных переменных для нейросетевой модели на основе анализа их влияния на выход. Математической основой такой процедуры выступает анализ конечных изменений, основанный на применении теоремы Лагранжа о промежуточной точке. Прикладная проблема исследования — выявление аномалий в наборах данных, полученных из информационной системы фиксации результатов оказания медицинской помощи.



Received 30.01.20