УДК 004.042

DEVELOPMENT OF A METHOD FOR USING AUTOENCODER TO SEARCH FOR ANOMALIES IN CLOUD DATA

C.S. Dzik ***I.I. Piletskii*** ***T.A. Asipovich***

C.S.Dzik

Graduate student BSUIR, Utech Solutions, software and data engineer, conduct scientific research of anomaly detection using autoencoder artificial neural network

I.I. Piletski

PhD, Associate Professor of the department of Informatics Department of Belarusian State University of Informatics and Radioelectronics. In the field of IT for over 50 years. Participation in the development of several dozen large projects: chief designer of the project, chief architect of software and information support, project manager, head of department.

T.A. Asipovich

Graduated from the Belarusian State Economic University. The area of scientific interests is related to the study of problems of e-commerce and entrepreneurship in the field of information technology.

Annotation. A methodology for conducting an experiment to search for anomalies in a cloud data array using an autoencoder is proposed. The technique was developed using an example and for use in analyzing the telemetry results of a solar power plant to search for defects and anomalies in its operation. However, with minor adaptation, it can be used to search for anomalies in other types of cloud data.

Keywords: autoencoder, PV Modules, cloud data, anomaly search, data defects.

Introduction. According to data quality experts, data is of high quality when it satisfies the requirements of its intended use. In other words, companies know that they have good quality data when they are able to use it to communicate effectively with their constituents, determine clients' needs, and find effective ways to serve their client base [1-4].

This data quality definition is broad enough to help companies with varying products, markets, and missions to understand if their data is up to standards. Data quality is not good or bad, high or low. It is a range or an indicator of operability of the data that passes through a company. Data quality management ensures the context-dependent process of improvement of suitability of the data, which is used for analysis and decision-making. The goal is to provide the vision of the «health» of the data by applying different processes and technologies to the increasingly complex data sets [1-4].

We want to be sure that when we take advantage of the cloud to help data managing, we define data quality parameters at the same time. The most obvious and compelling way to achieve the goal is to make sure we perform automatic data quality checks for all our data, wherever they are - in the cloud or elsewhere. We must always perform an on-site data quality check.

Virtually every company that works with data has a certain data quality (DQ) monitoring system. Some companies even hire an entire department that deals with the issue. This option is very expensive. In addition, most data quality checks are hard-coded and rule-based. In the event of a failure, the system notifies you of the risk indicator. Such rules are often critical to business continuity. For example, we cannot have a missing customer ID or a «risk profile» variable with an incorrect value. As the amount of data grows, you cannot specify a rule for work with each attribute; not to mention the difficulty of working with hard-coded multidimensional control checks.

The best option is automated DQ (data quality) checks using Machine Learning to detect anomalies that we don't even need to explicitly program.

For the DAD task the normal data easily reconstructed by the autoencoder, while the anomalous object for the model will be difficult to reconstruct. An autoencoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner.

Autoencoder are can: accept an input set of data; internally compress the data into a latent-space (low dimensional) representation; reconstruct the input data from the latent representation.

To accomplish this task, an autoencoder uses two components: an encoder and a decoder.

The encoder accepts the input data and compresses it into the latent-space representation. The decoder then attempts to reconstruct the input data from the latent space [5].

The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction. Along with the reduction side, a reconstructing side is learnt, where the autoencoder tries to generate from the reduced encoding a representation as close as possible to its original input [5 – 9].

The purpose of this work is to develop an experimental methodology for analyzing a large array of data using the example of telemetry results of solar power plants.

Development of the methodology. Figure 1 shows the general diagram of an autoencoder-type neural network. When analyzing big data, the autoencoder input for training must receive input data vectors in the form of parameter values that do not have defects or anomalies. That is, average statistical data that will depend on the problem being solved. After internal transformations in the autoencoder, its output must contain vectors of output data identical to the input. After training, when data vectors with defects or anomalies are fed to the input of the autoencoder, vectors without defects will be obtained at its output. By comparing the input and output vectors with the data, it will be possible to draw a conclusion about the presence of anomalous parameters in the input data vector.

Thus, to conduct an experiment to search for anomalies in the telemetry results of a solar power plant, it is necessary to develop an internal model of the autoencoder, develop a structure of input and output vectors for training and operation of the autoencoder, prepare a dataset for training the autoencoder, develop a mechanism and software for comparing the result of the autoencoder and conduct experiment on empirical data.

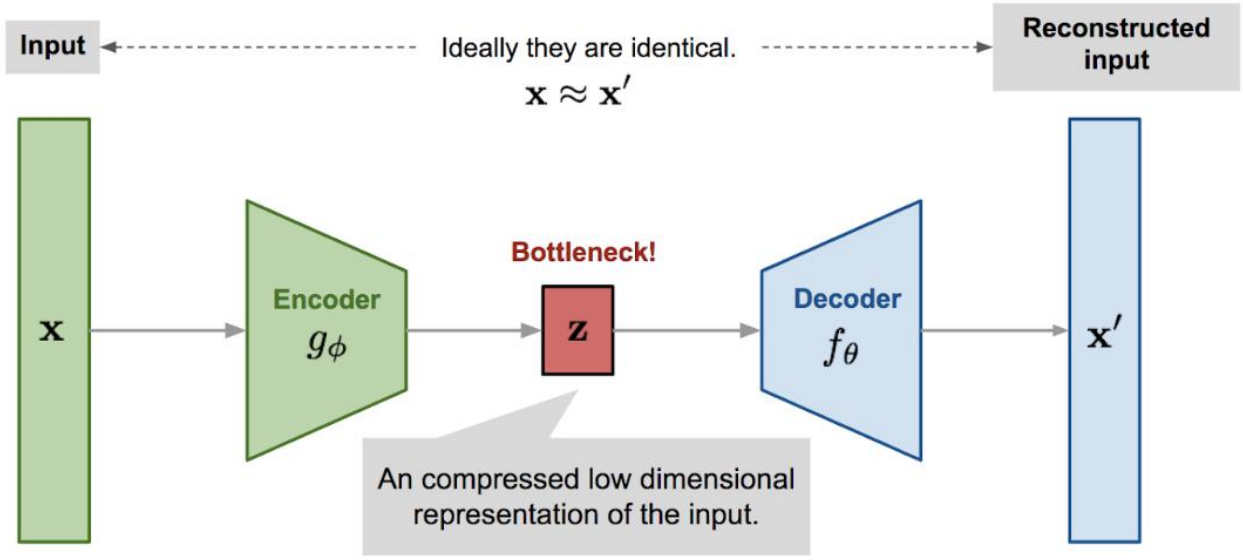

Figure 1. Example autoencoder neural network

Taking into account the specifics of the telemetry results of the operating parameters of the PV Modules of the power plant, we will use a vector of dimension 200 x 5 as a unit vector of input data, which will include the telemetry results from 10.00 to 18.00 hours of one day of the following parameters: timestamp, voltage, current, temperature in the PV Module housing, the level of illumination. The test data vector had a dimension of 200 x 3 and included the following parameters: timestamp, temperature in the PV Module housing, illumination level. At the output of the autoencoder, in both cases, a vector of 200 x 2 was obtained, that is, when test data is supplied to the input, the autoencoder must, based on the training results, restore the voltage and current values based on the temperature in the PV Module housing and the illumination level. In addition, to train the neural network, telemetry data of stably operating PV Modules (without defects and reduced efficiency) were used, selected as a result of direct analysis and the use of other methods of searching for anomalies in the operation of PV Modules.

To exclude telemetry results taken on days with partly cloudy and cloudy days from the training dataset, you need to use a filter that will leave data only on clear sunny days with low clouds. The filtering parameters are selected empirically and have the following values: current – 0 – 15 A, illumination – 360 – 1500 W/m^2, derivative with respect to current – -0.2 – 0.2, derivative with respect to illumination – -3.8 – 3, 8.

Figure 2 shows the voltage curves on Module 2.3_10 (defective) and Module 2.3_9 (good) during a sunny day. It can be seen that the voltage curves for these PV Modules are different. This is due to a malfunction in the Module 2.3_10: after the inverter entered the energy removal mode, the protective diode tripped, which disconnected a third of the cells from the circuit, so during the day the voltage on this PV Module is reduced.

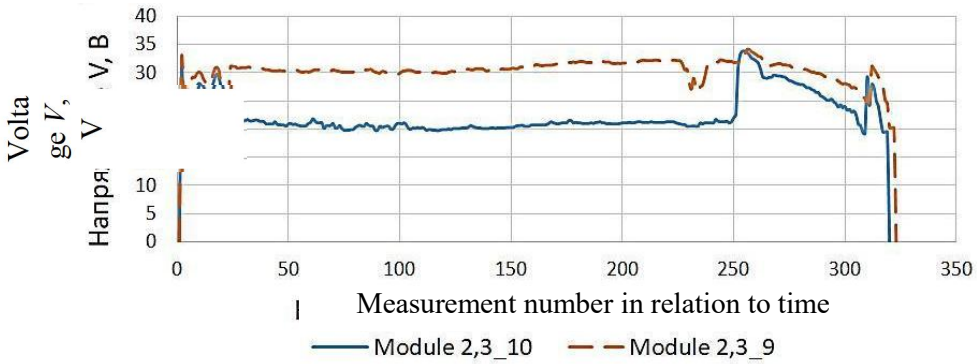

Figure 2. Change of voltage in Module 2,3_9 and Module 2,3_10 depending on time

Figure 3 shows the current curves on Module 2.3_1 (defective) and Module 1.5_12 (good) during a sunny day. The difference in the curves is due to the fact that in the cloud storage some of the points for the Module 2.3_1 PV Module are missing, and therefore it is shifted to the left. At the same time, the current curve for the PV Module Module 1.5_12 shows that at the beginning of the day there was cloudiness, which caused the telemetry collection device for the chain of PV Modules String 2.3 to be turned off. A defect in the data is the absence of some points on the current curve. Otherwise, the Module 2.3_1 works fine. The proposed filter will eliminate such situations due to the fact that points taken for the study will be taken during the absence of sharp changes in the readings of current strength and illumination.

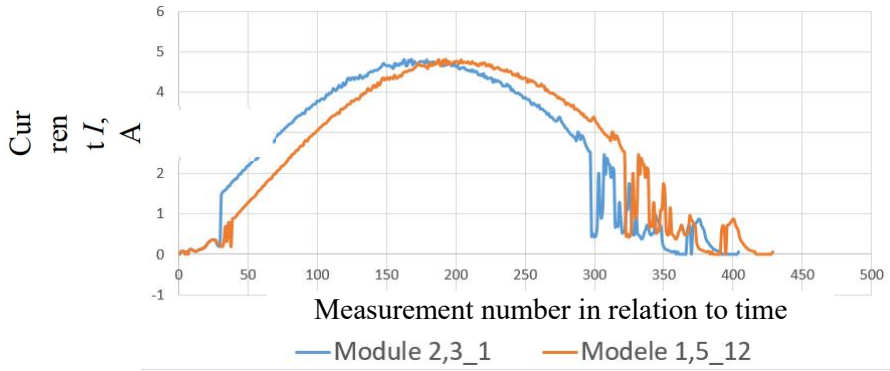

Figure 3. Change of current in Module 2,3_1 and Module 1,5_12 depending on time

The autoencoder developed to solve the problem consists of seven layers of neurons: input, output and five internal layers.

As a result of the autoencoder operation, current and voltage curves will be obtained for each PV Module depending on the illumination and temperature in the housing. These curves will be subject to comparison with curves obtained as a result of telemetry collection in automatic mode. A list of PV Modules with a curve divergence of more than 5% will be saved in a special file.

Conclusion. An experimental methodology, an autoencoder model, and a software algorithm for searching for anomalies in cloud data using the example of the results of collecting telemetry from solar power plants have been developed. The proposed approach will allow not only to search for anomalies in the operation of a solar power plant, but also to use it to search for anomalies in cloud data of a different nature and structure.

Reference list

[1] ИСО 8000-2 Качество данных. Часть 2. Словарь (ISO 8000-2, Data quality - Part 2: Vocabulary)

[2] ИСО/ТС 8000-110 Качество данных. Часть 110. Основные данные. Обмен данными характеристик. Синтаксис, семантическое кодирование и соответствие спецификации данных (ISO 8000-110, Data quality - Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification

[3] ИСО/ТС 8000-120 Качество данных. Часть 120. Основные данные. Обмен данными характеристик. Происхождение (ISO/TS 8000-120:2009, Data quality - Part 120: Master data: Exchange of characteristic data: Provenance)

[4] Data quality //[Online Resource] – Access Mode: https://en.wikipedia.org/wiki/Data_quality // Access Date: 14.02.2022

[5] U.A. Vishniakou, O.S. Koval, M.G. Mozdurani Shiraz. Use of Neural Networks for Detection and Recognition of the Anomalies in Enterprise Corporative Information System. Doklady BGUIR, 86 (4), 86-92.

[6] Alexander Prosak, Amitava Gangopadhyay and Hemant Garg A New Machine Learning Approach for Anomaly Detection Using Metadata for Model Training. EasyChair Preprint No 829 https://easychair.org/publications/preprint/Sf34

[7] Zhou, Chong, and Randy C. Paffenroth. "Anomaly detection with robust deep autoencoders." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017.

[8] Di Mattia, Federico, et al. "A Survey on GANs for Anomaly Detection." arXiv preprint arXiv:1906.11632 (2019).

[9] Malhotra, Pankaj, et al. "LSTM-based encoder-decoder for multi-sensor anomaly detection." arXiv preprint arXiv:1607.00148 (2016).

Авторский вклад

Константин Сергеевич Дик – разработка архитектурв автоэнкодера, программного обеспечения фильтрации и сравнения данных телеметрии, оформление результатов работы.

Иван Иванович Пилецкий – постановка задачи исследования, проработка плана работ, анализ полученных результатов.

Татьяна Анатольевна Осипович – анализ результатов телеметрии, разработка способов фильтрации данных телеметрии и сравнения их с результатами работы автоэнкодера.

РАЗРАБОТКА МЕТОДА ИСПОЛЬЗОВАНИЯ АВТОЭНКОДЕРА ДЛЯ ПОИСКА АНОМАЛИЙ В ОБЛАЧНЫХ ДАННЫХ

Дик К.С.
Аспирант БГУИР, Ютех Солюшнс, инженер по программному обеспечению и данным

И.И. Пилецкий
к.ф.-м.н., доцент кафедры информатики Белорусского государственного университета информатики и радиоэлектроники. В сфере IT более 50 лет. Участие в разработке нескольких десятков крупных проектов

Т.А. Осипович
Доцент кафедры экономики БГУИР, кандидат экономических наук

Аннотация. Предложена методика проведения эксперимента по поиску аномалий в массиве облачных данных с использованием автоэнкодера. Методика разработана на примере и для использования при анализе результатов телеметрии солнечной электростанции на предмет поиска дефектов и аномалий в её работе. Однако, при незначительной адаптации, может быть использована при поиске аномалий в других видах облачных данных.

Ключевые слова: автоэнкодер, солнечные панели, облачные данные, поиск аномалий, дефекты данных.