DEVELOPMENT OF METHODS AND TOOLS FOR ANOMALY DETECTION IN DATA ANALYSIS TASKS

Dmytro Palahin

Project Objective

The goal of the project is to increase the efficiency of machine learning algorithms by detecting and eliminating anomalies in data when applying new proposed processing methods in the tasks of analyzing intrusions into computer networks, detecting financial fraud, diagnostic analysis, etc..

Object and subject of research

The object of research is the process of anomaly detection in information and telecommunication systems for solving machine learning problems.

The subject of the research is machine learning methods for anomaly detection in data, software tools for algorithmic implementation of anomaly detection processes.

The novelty of the project

- A new author's method of improving the quality of anomaly detection in data based on a neural network structure has been developed
- A method for iterative data cleaning and a method for creating heat maps have been developed
- A new method of implementing semi-automatic learning is proposed, which significantly reduces resources for data preparation.
- As test tasks, model data structures and a dataset of banking operations with anomalous data have been proposed.

Research methods

► Synthesis of the neural network structure of data analysis

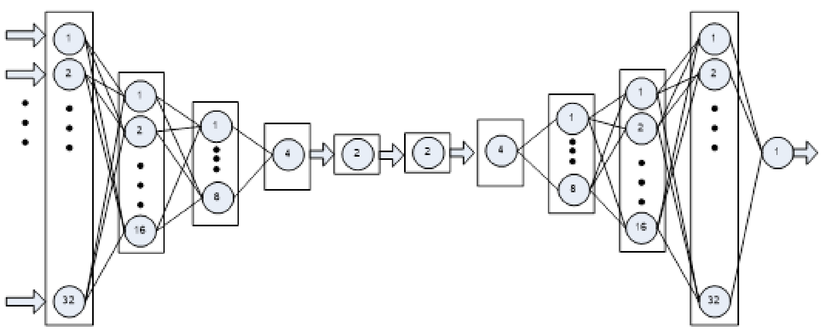


Fig.1. Presentation of the Dense Layer autoencoder model - "**Autoencoder Classifier**".

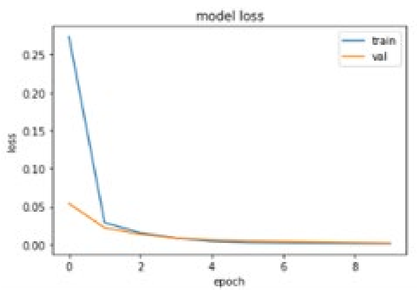


Fig.2. Presentation of the learning process of the neural network - "**Autoencoder Classifier**"

► A Method of Creating «Heat Maps» for Data Analysis

Selection of data processing methods → **Creating a histogram of the results of "voting" methods** → **Histogram analysis and data cleaning threshold selection** → **Separation of normal and abnormal data**

► Semi-Automatic Anomaly Detection Method

Training a classifier to determine the probability that a sample is labeled

P(m=1)

Applying a classifier to determine the probability that a positive sample is labeled,

P(m=1| y=1)

Application of a classifier to determine the probability of labeling a sample,

P(m=1)

Estimating the probability that the sample v is positive,

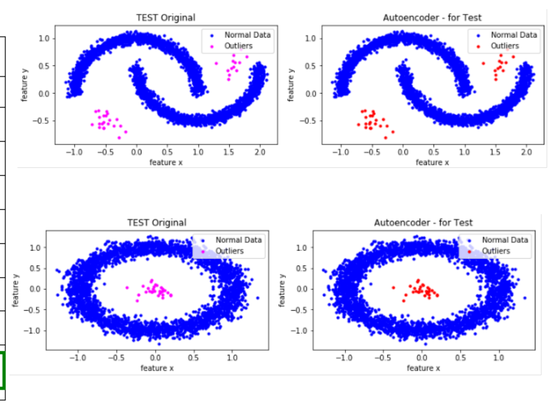
P(y=1) = P (m=1) / P (m=1|y=1)

Research results

Indicators of the effectiveness of fraud detection (Credit-card Fraud) in financial transactions based on data from the Kaggle platform

№	Method name	Performance indicator				
		f1	roc_auc	accuracy	precision	recall
1	KNN	0.705	0.864	0.991	0.643	0.864
2	LOF	0.551	0.576	0.986	0.539	0.576
3	CBLOF	0.653	0.737	0.989	0.614	0.737
4	XGBOD	0.906	0.854	0.998	0.977	0.854
5	AutoEncoder	0.651	0.721	0.989	0.615	0.721
6	*AdaBoostClassifier*	0.903	0.870	0.998	0.941	0.870
7	KNeighborsClassifier	0.891	0.851	0.988	0.922	0.841
NEW result	Autoencoder Classifier	**0.925**	**0.870**	**0.998**	**0.999**	**0.870**

The result of data anomaly detection



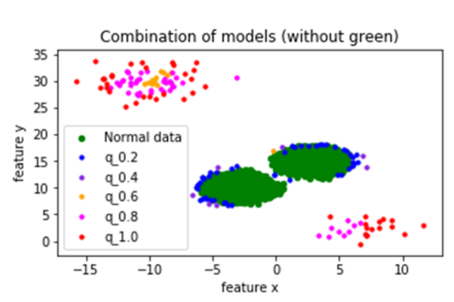


Fig.3. Heat map of 5-fold differentiation of data by basic methods

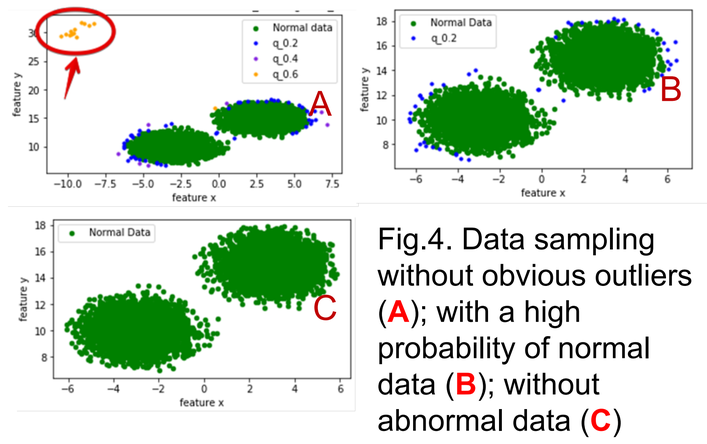
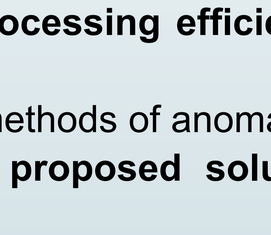
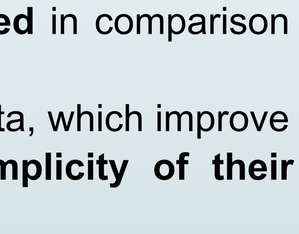


Fig.4. Data sampling without obvious outliers (**A**); with a high probability of normal data (**B**); without abnormal data (**C**)

The Effectiveness of Anomaly Detection

Partial data labeling	Partial labeling of positive data
KNeighborsClassifier	
	
RandomForestClassifier	

Results and conclusions

1. A **new author's method** of improving the quality of abnormal data detection has been developed, which is based on the application of deep learning technology and the development of a neural network structure, which, on model examples, demonstrates an increase in the accuracy of anomaly detection compared to basic data processing methods.
2. A method of **iterative data cleaning and creation of "heat maps" of data** based on their probability distribution has been developed, which makes it possible to obtain arrays with a given probability of normal data.
3. A new **semi-automatic learning method** is proposed, which allows using not the entire set of labeled data for solving the AD problem, but only the part that satisfies the given accuracy for anomaly detection, which significantly reduces both time and resources in data preparation.
4. The proposed methods were used to clean the data from outliers, and the **data processing efficiency was analyzed** in comparison with other known methods using quantitative and qualitative characteristics..
5. The implementation of model experiments confirmed the possibility of applying new methods of anomaly detection in data, which improve the quality of data preparation in comparison with known results. The **advantage of the proposed solutions is the simplicity of their practical implementation and the quality of data processing**.