ABSTRACT

Master dissertation: 102 p., 40 fig., 1 tab., 2 sup., 62 sources.

Relevance: machine learning methods are used where conventional algorithms cannot be applied due to the complexity of the problem and the impossibility of solving it by traditional methods. However, the amount of data needed for learning is constantly growing and increasingly cannot be processed quickly and efficiently by a single work device. The solution to this problem is the use of distributed computing and the application of such approaches to machine learning problems using distributed systems with multiple computing nodes and network interaction between them. Distribution can not only speed up learning, but also increase bandwidth, use data streams, perform optimizations on models, teach different versions in parallel, and more.

Purpose: an acceleration of machine learning due to the method of distributed machine learning on the example of solving the problem of finding anomalies using isolation trees.

To achieve this goal, the following tasks were formulated:

– perform an analysis of existing methods and approaches to distributed machine learning;

– collection of training data and formation of sets for distribution;

– to develop a method of distributed machine learning on the example of the isolation tree algorithm;

– testing and analysis of the effectiveness of the obtained method;

– determining the further direction of research.

Object of study: processes of distributed machine learning.

Subject of study: methods of distributed machine learning.

Research methods: isolation forest and trees, distributed computing, GFS file system, MapReduce computational approach, data flows were used to solve this problem.

Scientific novelty: The scientific result of the master's dissertation is the creation of a method of distributed learning based on the use of distributed data, computing resources and the involvement of streaming data processing.

The practical value: is determined by the fact that the proposed method allows to accelerate the learning of models using isolation trees, increase the fault tolerance of the system, and maintain transparent scalability for the user.

Relationship with working with scientific programs, plans, topics: work was performed at the Department of Automated Information Processing and Management Systems of the Igor Sikorsky National Technical University of Ukraine «Kyiv Polytechnic Institute» within the topic «Methods and technologies of high-performance computing and processing of large data sets». State registration number 0117U000924.

Approbation: The main provisions of the work were reported and discussed at the IV All-Ukrainian Scientific and Practical Conference of Young Scientists and Students «Information Systems and Management Technologies» (ISTU-2020), as well as at the XVI International Scientific Conference «Intellectual Systems of Decision-making and Problem of Computational Intelligence» (ISDMCI'2020).

Keywords: MACHINE LEARNING, DISTRIBUTED LEARNING, ANOMALY DETECTION, STREAM PROCESSING, DATA FLOWS.