ABSTRACT

Relevance of the research We encounter significant increase of data in the global network and processing of that data becoming more and more difficult. There are a lot of machine learning and data classifications algorithms that come to the rescue and help automate manual work and analysis. Amount of users who want to have easier access to that data in unified format is also increases. That is why algorithms for data classification and clusterization have gain their growth. Speaking from the point of data scientiests these algorithms are not the best option in every use case and need to be tuned in order to achieve the best results. Therefore universal methods that will be able to automatically classify various input data have such a high demand on the market. Moreover those algorithms are required to have intuitive implementation not only for the creator of the algorithm but for the users and regular developers too.

Object of research is a process of creating an algorithm to build universal predictive model.

Subject of research is a set of methods for building predictive models and algorithms of data classification.

Project's goal: developing a new algorithm for predictive model creation that will show the same accuracy as competitors and have simple implementation.

Methods of research. There are methods of data mining, data classification techniques and statistical methods in current project.

Scientific innovation of the projects includes:

1. Universality of a model that allows to eliminate differences between internal implementations of algorithms and use it with the same data not losing the accuracy among with better performance results.

2. Process of creation for such a model was developed and generialized for any arbitrary one.

3. Permormance of the model develped was confirmed.

On the current state of the projects this approach allows usage of a model on real-world data and might be used as drop-in replacement for existing algorithms.

Practical value of the projects is based on the significant increase of the performance when using the model created by the method described. Using model approximation and similar set of cpu instructions can outcome in performance boost. The other advantage allows to decrease average time for data scientiests to spend while examining internal structure of the models and details of algorithm's implementation. Model shows great results on a large datasets too and therefore it will decrease total amount of time when launching multiple times or using it with large data.

Project's approbation. Main ideas and summary for the projects were presented and discussed during The IX annual scientific conference "Applied math and computing" and has been published in corresponding set of theses. Data mining and its preprocessing as well as publishing of the results has been made on the web-resource kpidata.org; access to the input data is free and data is stored under version control system on the GitHub (github.com/kpidata/datasets).

Project's structure. Master's dissertation consists of introduction, five sections, summary and appendix.

Introduction shows general overview of the project, description of the current state for the problem and shows the relevance of such a topic of research.

First section describes theoretical background, current set of algorightms that are used for text classification, mathematical background for those methods and algorithms for predictive models creation. General approaches that are used for text classification have been overviewed. Main point of the section is a process of data mining starting from initial collection of data and resulting in usage of predictive model.

Second section contains analysis of existing algorithms and research of their

internal implementation was made. Advantages and disadvantages of each class of algorithms was showed and fields for their application were described. List of requirements for the resulting algorithm was created and then suggested a ways for both optimization and improvement. Influence of the changes on the resulting model was inspected and then necessity of making these changes was confirmed.

Third section shows implementation of every stage for the method described; architecture and design solution for the resulting software are discussed; usage of microservices approach is justified; structure and internal implementation of each microservice is provided as well as charts and diagrams that shows interaction between components of the system.

Forth section provides results of algorithm usage and significant performance increase was confirmed. Confirmation for the advantages of usage of homogeneous cpu instructions has been made. The comparison of accuracy and performance for the developed algorithm and its competitors was conducted. Ways for the future development and improvement were suggested.

In fifth section analysis of software is provided as well as general estimation and possibilities to join a market. Strong and weak sides of the project are highlighted and comparison with competitors is made. Estimation of investments amount and the amount of resources that need to be invoked for the successful results has been made.

Summary briefly overviews achieved results and highlights key features of the work done.

Appendix consists of significant source code snippets and code for the main modules.

The project consists of 92 pages, has 1 additional code listing, references with 20 entries, 14 figures and 24 tables.

Keywords: classification, predictive models, model approximation, dataset, machine learning.