

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

КАФЕДРА ИНФОРМАЦИОННЫХ СИСТЕМ УПРАВЛЕНИЯ

**АННОТАЦИЯ К ДИПЛОМНОЙ РАБОТЕ
«АЛГОРИТМЫ ОБРАБОТКИ НЕСБАЛАНСИРОВАННЫХ ДАННЫХ»**

Клицунова Екатерина

Научный руководитель – кандидат технических наук,
доцент ИСУ ФПМИ
Лукашевич М. М.

Минск, 2025

АННОТАЦИЯ

Дипломная работа, 49 страниц, 8 таблиц, 16 иллюстраций, 4 формулы, 5 приложений, 26 источников.

Ключевые слова: НЕСБАЛАНСИРОВАННЫЕ ДАННЫЕ, МЕТОДЫ БАЛАНСИРОВКИ ДАННЫХ, МАШИННОЕ ОБУЧЕНИЕ, КЛАССИФИКАЦИЯ, ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ, РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА, ВЕБ-ПРИЛОЖЕНИЕ

Объект исследования – алгоритмы и методы машинного обучения, предназначенные для работы с несбалансированными наборами данных.

Предмет исследования – влияние методов балансировки данных на качество моделей классификации при несбалансированности данных.

Цель исследования – провести экспериментальное исследование влияния методов балансировки данных на эффективность классических моделей классификации. Разработать приложение, рекомендуемое алгоритмы балансировки в зависимости от размера набора данных и величины дисбаланса, и дающее возможность провести балансировку данных.

Методы исследования: сравнительный анализ алгоритмов для работы с несбалансированными данными, постановка и реализация эксперимента, проектирование модели ранжирования для рекомендации алгоритмов балансировки данных, проектирование и разработка приложения.

Полученные результаты и их новизна: оценено влияние методов балансировки данных и подбора гиперпараметров на качество модели в задачах классификации. Разработано веб-приложение с моделью рекомендаций методов балансировки. Новизна работы заключается в интеграции алгоритма подбора методов балансировки с практической реализацией программного инструмента, позволяющего применять стандартные методы балансировки данных.

Достоверность материалов и результатов дипломной работы. Автор работы подтверждает, что приведенный в ней расчетно-аналитический материал правильно и объективно отражает состояние объекта исследования. Все заимствованные из источников теоретические, методологические и методические положения и концепции сопровождаются ссылками на их авторов.

Область практического применения: оптимизация выбора подходящих методов работы с несбалансированными данными для конкретных практических задач в сферах, где данная проблема является критической: медицине, биоинформатике, инженерии, безопасности, бизнесе и других.

АНАТАЦЫЯ

Дыпломная праца, 49 старонак, 8 табліц, 16 ілюстрацый, 4 формулы, 5 дадаткаў, 26 крыніц.

Ключавыя словы: НЕБАЛАНСІРОВАННЫЯ ДАДЗЕННЫЯ, МЕТАДЫ БАЛАНСІРОЎКІ ДАДЗЕННЫХ, МАШЫНАЕ НАВУЧАННЕ, КЛАСІФІКАЦЫЯ, ЭКСПЕРЫМЕНТАЛЬНАЕ ДАСЛЕДАВАННЕ, РЭКАМЕНДАЎНАЯ СІСТЭМА, ВЭБ-ДАДАТАК

Аб'ект даследавання – алгарытмы і метады машыннага навучання, прызначаныя для працы з незбалансаванымі наборамі дадзеных.

Прадмет даследавання – уплыў метадаў балансавання дадзеных на якасць мадэляў класіфікацыі пры незбалансаванасці дадзеных.

Мэта даследавання – правесці эксперыментальнае даследаванне ўплыву метадаў балансавання дадзеных на эфектыўнасць класічных мадэляў класіфікацыі. Распрацаваць дадатак, які рэкамендуе алгарытмы балансавання ў залежнасці ад памеру набору дадзеных і велічыні дысбалансу, і дае магчымасць правесці балансаванне дадзеных.

Методы даследавання: параўнальны аналіз алгарытмаў для апрацоўкі незбалансаваных дадзеных, пастаноўка і рэалізацыя эксперыменту, праектаванне мадэлі ранжыравання для рэкамендацыі алгарытмаў балансавання дадзеных, праектаванне і распрацоўка прыкладання.

Атрыманыя вынікі і іх навізна: ацэнены ўплыў метадаў балансавання даных і падбору гіперпараметраў на якасць мадэлі ў задачах класіфікацыі. Распрацавана вэб-дадатак з мадэллю рэкамендацый метадаў балансавання. Навізна працы складаецца ў інтэграцыі алгарытму падбору метадаў балансавання з практычнай рэалізацыяй праграмнай прылады, які дазваляе ўжываць стандартныя метады балансавання дадзеных.

Дакладнасць матэрыялаў і вынікаў дыпломнай працы. Аўтар працы пацвярджае, што прыведзены ў ёй разлікова-аналітычны матэрыял правільна і аб'ектыўна адлюстроўвае стан аб'екта даследавання. Усе запазычаныя з крыніц тэрэтычныя, метадалагічныя і метадычныя палажэнні і канцэпцыі суправаджаюцца спасылкамі на іх аўтараў.

Вобласць магчымага практычнага прымянення: аптымізацыя выбару падыходных метадаў работы з незбалансаванымі дадзенымі для канкрэтных практычных задач у сферах, дзе дадзеная праблема з'яўляецца крытычнай: медыцыне, біяінфарматыцы, інжынерыі, бяспекі, бізнесе і іншых.

ANNOTATION

Diploma work, 49 pages, 8 tables, 16 illustrations, 4 formulas, 5 appendixes, 26 references.

Keywords: IMBALANCED DATA, DATA BALANCING METHODS, MACHINE LEARNING, CLASSIFICATION, EXPERIMENTAL STUDY, RECOMMENDATION SYSTEM, WEB APPLICATION

The object of the research is the algorithms and methods of machine learning designed to work with imbalanced datasets.

The subject of the research is the influence of data balancing methods on the quality of classification models trained with imbalanced datasets.

The aim of the research is to conduct an experimental study of the influence of data balancing methods on the quality of classical classification models, and to develop a software tool that recommends balancing algorithms depending on the size of the dataset and the imbalance ratio, while also enabling data balancing.

Research methods. Comparative analysis of algorithms for working with imbalanced data, setup and implementation of the experiment, results interpretation, design of a machine learning model for recommending data balancing algorithms, design and development of an application.

The results of the work and their novelty. The study assessed the impact of data balancing methods and hyperparameter tuning on the quality of classification models. A web application with a model, which recommends balancing methods, has been developed. The novelty of the work lies in integrating an algorithm for selecting balancing methods with the practical implementation of a software tool that allows the use of standard data balancing methods.

Authenticity of the materials and results of the diploma work. The author confirms that the analytical and computational material presented in this work it correctly and objectively reflects the state of the research object. All theoretical, methodological and methodological provisions and concepts borrowed from literary and other sources are accompanied by references to their authors.

Recommendations on the usage. The results of the work can be used in optimization of the process of selecting suitable methods and algorithms for working with imbalanced data for specific practical problems in areas where this problem is critical such as medicine, bioinformatics, engineering, security, business, and others.