АНОТАЦІЯ

Мілян Н. В. Моделі і методи прийняття рішень при створенні клінічних діагностичних систем. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки». – Тернопільський національний технічний університет імені Івана Пулюя, Тернопіль, 2021.

Зміст анотації. Дисертація присвячена вирішенню актуального наукового завдання розвитку методів класичного, глибокого та наукового машинного навчання в умовах невизначеності з використанням мінімаксного підходу.

У вступі обґрунтовано актуальність дослідження, наведено зв'язок роботи з науково-дослідними темами, визначено мету та завдання дослідження, об'єкт та предмет дослідження, наведено перелік методів дослідження, що застосовувались для досягнення мети дисертаційної роботи. Сформульовано наукову новизну, практичне значення отриманих результатів та особистий творчий внесок здобувача. Подано відомості щодо апробації та опублікування результатів дослідження.

У першому розділі «Математичне та програмне забезпечення інформаційних технологій задач машинного навчання в умовах невизначеності (огляд літератури)» проаналізовано основні області використання і можливості Big Data. Розглянуто методики аналізу великих даних. Відзначено, що значної уваги заслуговує аналіз даних у системних медичних дослідженях. У зв'язку із зростання обсягу даних, також розвиваються технології аналітики великих даних на основі машинного навчання.

Розглянуто технологічні питання задач машинного навчання на основі хмарних обчислень. Здійснено загальний огляд хмарних платформ та розглянуто парадигми хмарних обчислень. Проаналізовано сервіси хмарних платформ та необхідні дії з точки зору розгортання додатків. За основу взято

Google Cloud Platform та його сервіс Google App Engine, який дозволяє розробникам створювати масштабовані додатки на Python і Java, а також володіє усім необхідним для реалізації алгоритму індукції дерева рішень C5.0.

Розглянуто алгоритмічні питання задач машинного навчання в умовах невизначеності, зокрема розглянуто такі типові алгоритми прийняття рішень: дерево рішень, випадковий ліс, алгоритм k-середніх, метод опорних векторів, наївний баєсівський класифікатор, ансамблеві методи.

Наведено типові види невизначеностей у медичних застосуваннях: алеаторна та епістемічна, а також дано визначення поняття наукового машинного навчання.

Проведено огляд результатів в галузі машинного навчання на основі мінімаксного підходу. Обгрунтовано потребу розробки методів на основі мінімаксного підходу до розв’язування задач класичного та глибокого машинного навчання.

На основі проведеного аналізу визначено наукове завдання, яке полягає у розробці програмного та математичного забезпечення інформаційних технологій, пов’язаних із прийняттям рішень у системних медичних дослідженнях, а саме: реалізація алгоритмів машинного навчання на основі хмарних обчислень, розробка методів на основі мінімаксного підходу до розв’язування задач класичного та глибокого машинного навчання, розвиток методів наукового машинного навчання в кіберфізичних системах медико-біологічних процесів.

У другому розділі «Технологічні питання реалізації алгоритмів машинного навчання на основі хмарних обчислень у випадку системних медичних досліджень» здійснено математичне представлення задачі індукції дерева рішень для реалізації на хмарних платформах. Наведено алгоритм для генерації дерева рішень.

Запропоновано приклад реалізації індукції дерева рішень, що є прототипом для реалізації на основі хмарної платформи, для системних

медичних досліджень. Метод продемонстровано на прикладі діагностування серцевої недостатності на основі даних 63-х пацієнтів навчально-практичного центру первинної медико-санітарної допомоги Тернопільського національного медичного університету імені І.Я.Горбачевського.

Розглянуто питання архітектури систем прийняття рішень на основі системних медичних досліджень. Інформаційні системи для медичної практики та досліджень здебільшого грунтуються на монолітній архітектурі. У свою чергу мікросервіс – це інша парадигма, яка служить для організації та використання розподілених послуг.

Здійснено концептуальне представлення інформаційної моделі задачі прийняття рішень для невідкладної медичної допомоги при настанні серцевої недостатності та представлено підхід до реалізації рішення задачі невідкладної медичної допомоги на основі архітектури мікросервісів, підхід до реалізації базується на використанні фреймворків Spring, Spring Boot та Spring Cloud frameworks.

У третьому розділі «Задачі класичного машинного навчання в умовах невизначеності інтегровані з даними із систем з відкритим кодом електронних медичних записів» розглянуто задачу класичного машинного навчання інтегровану з даними із систем з відкритим кодом електронних медичних записів в умовах визначеності. А саме: здійснено постановку задачі машинного навчання, інтегровану з даними з систем з відкрити кодом електронних медичних записів; наведено математичний опис задачі машинного навчання в медичних дослідженнях в умовах визначеності; розроблено модель машинного навчання для медичних досліджень; розглянуто та розв’язано задачу імпорту даних з систем з відкритим кодом електронних медичних записів; розглянуто задачі зменшення розмірності даних та класифікації; наведено результати експериментальних досліджень, що пов’язані з прогнозуванням переломів для жінок середнього віку, що є на сьогодні актуальним завданням профілактичної медицини.

У процесі експериментальних досліджень отримано моделі машинного навчання на основі п’яти методів. Результати ефективності методів оцінено за допомогою середньої помилки неправильної класифікації. Досліджено, що найточнішою моделлю, відповідно до оцінки якості прогнозування, є метод випадкового лісу. Дерево рішень індукується на основі атрибутів після зменшення розмірності.

Розглянуто задачу класичного машинного навчання в умовах епістемічної та алеаторної невизначеності. А саме: здійснено математичний опис задачі машинного навчання в умовах невизначеності; розглянуто питання зменшення розмірності даних в умовах невизначеності, запропоновано модифікацію методу головних компонент щодо стратегії передискретизації; представлено загальну блок-схему побудови моделі машинного навчання на основі мінімаксного критерію в умовах невизначеності.

Реалізовано на основі розробленої схеми машинного навчання приклад для кардіологічних даних. Експериментальне дослідження включає дані 1651 пацієнта з діагнозом інфаркт міокарда. Цільовим атрибутом прогнозування є тривалість життя. Дані кожного пацієнта включають 97 атрибутів, що містять як числові, так і категоріальні значення. Розглянуто набір Ψ, що включає модель лінійної регресії, модель на основі методу опорних векторів та модель випадкового лісу. У випадку середньоквадратичного відхилення, як показника ефективності, модель на основі методу опорних векторів є розв’язком задачі машинного навчання на основі мінімаксного підходу. У випадку часу навчання, як показника ефективності, модель лінійної регресії є розв’язком задачі машинного навчання на основі мінімаксного підходу.

У четвертому розділі «Глибоке машинне навчання в умовах невизначеності на основі мінімаксного підходу та розробка моделі SciML» пропонується реалізація архітектури нейронної мережі на основі щільної U-Net мережі у середовищі Google Collab на загальнодоступному наборі даних DRIVE з використанням бібліотеки машинного навчання Tensorflow для

сегментації судин сітківки очного дна. Представлений загальний процес роботи нейромережі.

Для визначення помилки між реальними і отриманими результатами та її мінімізації в якості функції втрат використано бінарну крос-ентропію. Для оцінки методу використано точність, повноту, матрицю помилок, площу під кривою помилок. Для нормалізації зображення було поділено на плитки розміром 8x8, 16x16, 25x25, 40x40.

У випадку бінарної крос-ентропії як показника ефективності, модель U-Net з плитками 8x8 є розв’язком задачі машинного навчання на основі мінімаксного підходу. На першому кроці порівнюється значення функції втрат для всіх розглянутих моделей. На другому кроці визначено, що значення бінарної крос-ентропії для моделі U-Net з плитками 8x8 буде мінімальним серед максимальних характеристик.

У випадку з часом навчання, як показника ефективності, модель U-Net з плитками 25x25 є розв’язком задачі машинного навчання на основі мінімаксного підходу. На першому кроці порівнюється значення часу тренування всіх моделей, що розглядаються. На другому кроці визначено, що значення часу для моделі U-Net з плитками 25x25 буде мінімальним серед максимальних показників.

Розглядається кіберфізична модель для біосенсорних систем, що грунтується на математичному описі неперервної динаміки популяцій у поєднанні з динамічною логікою, яка використовується для дискретних подій. Дослідження стійкості грунтується на алгоритмі наукового машинного навчання, що пропонує розв’язування задачі класифікації кіберфічних біосенсорних систем. Алгоритм наукового машинного навчання включає в себе такі етапи: означення класів траєкторій систем, генерування матриці випадкових початкових значень запізнення часу і параметрів швидкості, запуск моделі і класифікація вхідних даних, побудова матриці зв’язків між початковими значеннями, запізненням в часі та між параметрами швидкості,

застосування індукції дерева рішень алгоритму C5.0 до взаємозв'язку між початковими значеннями і параметрами швидкості.

Спосіб дозволяє оптимальним чином підібрати параметри кіберфізичної біосенсорної системи (початкові умови, швидкісні параметри та час імунної відповіді).

Основні наукові результати дисертації опубліковано у 11 працях, зокрема: дві статті [1], [3] у наукових фахових періодичних виданнях України; одна стаття у закордонному фаховому періодичному виданні [2]; видано один патент на корисну модель [11]; 7 публікацій [4], [5], [6], [7], [8], [9], [10] у матеріалах міжнародних та всеукраїнських наукових, науково-технічних конференцій. З них одна робота входить до міжнародної наукометричної бази Scopus [4], дві роботи [1], [3] входять до міжнародної наукометричної бази Index Copernicus.

Ключові слова: машинне навчання, прийняття рішень в умовах невизначеності, системні медичні дослідження, глибоке навчання, наукове машинне навчання, мінімаксний підхід, дерево рішень, хмарні обчислення.

SUMMARY

Milian N. V. Models and methods of decision making in the creation of clinical diagnostic systems. - Qualifying scientific work on the rights of the manuscript.

Thesis for the degree of Doctor of Philosophy in the specialty 122 «Computer Science». - Ternopil Ivan Puluj National Technical University, Ternopil, 2021.

Contents of the summary. The thesis is devoted to the solving of a relevant scientific problem of the development of methods of classical, deep, and scientific machine learning under uncertainties with the use of the minimax approach.

The introduction explains the relevance of the study, discovers the connection between the work and the research topics, sets the purpose and defines the research object and subject, lists the research methods used to achieve the goal of the thesis. The scientific novelty, practical significance of the obtained results, and personal creative contribution of the applicant are formulated. Information on approbation and publication of research results is given.

The first section «Mathematical and software information technology of machine learning problems under uncertainties (literature review)» analyzes the main areas of use and capabilities of Big Data. Methods of Big Data analysis are reviewed. It is noted that the analysis of data in systemic medical research deserves considerable attention. It is observed that due to the growth of data, the technology of Big Data analytics based on machine learning is also developing.

Technological issues of machine learning problems based on cloud computing are considered. A general overview of cloud platforms is provided and paradigms of cloud computing are studied. Cloud platform services and necessary actions in terms of application deployment are considered. The work is based on the Google Cloud Platform and its Google App Engine service, which allows developers to create scalable applications in Python and Java, and also has everything necessary for the implementation of the decision tree C5.0 algorithm induction.

Algorithmic issues of machine learning problems under uncertainties are considered. In particular, the following decision-making algorithms are reviewed: a decision tree, a random forest, a k-means clustering, a support vector machine, naive Bayesian classifier, the ensemble methods.

Typical types of uncertainties in medical applications are given: aleatory and epistemic. The definition of scientific machine learning is given.

A review of results in the field of machine learning based on the minimax approach is conducted. The need to develop methods based on the minimax approach to solving problems of classical and deep machine learning is substantiated.

Based on the conducted analysis, the scientific task is set. It consists in the development of software and mathematical software of the information technologies related to decision-making in the system of medical researches, namely implementing the algorithms of machine learning based on cloud calculations, development of the minimax approach methods of solving problems of classical and deep machine learning, development of methods of scientific machine learning in cyber-physical systems of medical and biological processes.

The second section, «Technological issues of implementation of machine learning algorithms based on cloud computing in the case of systemic medical research», gives a mathematical presentation of the problem of induction of the decision tree for implementation on cloud platforms. An algorithm for generating a decision tree is presented.

An example of the implementation of the decision tree induction which is a prototype for implementation on the basis of a cloud platform for systematic medical research is given. The method is demonstrated on the example of diagnosing heart failure based on the performance of 63 patients of the practical center of primary health care of Ternopil National Medical University named after I. Gorbachevsky educational.

The issues of the architecture of decision-making systems based on systemic medical research are considered. Information systems for medical practice and

research are mostly based on monolithic architecture. Besides, microservices are another paradigm that serves to organize and use distributed services.

A conceptual presentation of the information model of the decision-making problem for emergency medical care in case of heart failure is made and the approach to the implementation of the solution of the emergency medical care problem based on microservice architecture is presented, the approach to implementation is based on Spring, Spring Boot, and Spring Cloud frameworks.

In the third section «Problems of classical machine learning under uncertainties are integrated with data from open-source systems of electronic medical records», the problem of classical machine learning integrated with data from open-source systems of electronic medical records under certainties is considered. Namely, the machine learning problem of electronic medical records integrated with data from open source systems was set; a mathematical description of the machine learning problem in medical research under certainties is given; the machine learning model for medical researches is developed; the problem of importing data from open-source systems of electronic medical records is considered and solved; the problems of dimensional reduction and classification are considered; the results of experimental studies related to the prediction of fractures for middle-aged women, which is currently an urgent task of preventive medicine.

Experimental studies have resulted in machine learning models based on five methods. The results of the effectiveness of the methods were evaluated using the average error of incorrect classification. It is investigated that the most accurate model according to the assessment of forecasting quality is the method of random forest. The decision tree is induced based on attributes after dimension reduction.

The problem of classical machine learning under epistemic and aleatory uncertainties is considered. Namely, a mathematical description of the problem of machine learning under uncertainties is provided; the issue of reducing the dimensionality under uncertainties is considered, the modification of the principal component analysis method on the resampling strategy is proposed; the general

block diagram of the machine learning model construction based on a minimax criterion under uncertainties is presented.

The example for cardiac data based on the developed scheme of machine learning is implemented. The experimental study includes data from 1651 patients diagnosed with myocardial infarction. The target attribute of forecasting is life expectancy. Each patient's data includes 97 attributes that contain both numerical and categorical values. We considered the set Ψ, which includes a linear regression model, a support vector machine, and a random forest model. In the case of standard deviation, as an indicator of efficiency, the support vector machine model the solution of the machine learning problem based on the minimax approach. In the case of training time as an indicator of efficiency, the linear regression model is the solution of the machine learning problem based on the minimax approach.

The fourth section, «Deep machine learning under uncertainties based on a minimax approach and the development of the SciML model» proposes implementing a neural network architecture based on a dense U-Net network in Google Collab on a public DRIVE data set using the Tensorflow, the machine learning library, for retinal vascular segmentation. The general process of the neural network is presented.

Binary cross-entropy was used as a function of losses to determine the error between the actual and obtained results and to minimize it. To evaluate the method we used the metrics - the precision, the recall, the confusion matrix, and the receiver operating characteristic curve. To normalize the image, it was divided into 8x8, 16x16, 25x25, 40x40 tiles.

In the case of binary cross-entropy as an indicator of efficiency, the U-Net model with 8x8 tiles is the solution of the machine learning problem based on the minimax approach. The first step compares the value of the loss function for all considered models. In the second step, it is determined that the value of binary cross-entropy for the U-Net model with 8x8 tiles will be the minimum among the maximum characteristics.

In the case of training time as an indicator of efficiency, the U-Net model with 25x25 tiles is the solution of the machine learning problem based on the minimax approach. In the first step, we compare the training time values of all the models that are studied. In the second step, it is determined that the time value for the U-Net model with 25x25 tiles will be the minimum among the maximum values.

A cyber-physical model for biosensor systems is considered, which is based on a mathematical description of the continuous dynamics of populations in combination with the dynamic logic used for discrete events. The study of stability is based on the algorithm of scientific machine learning, which offers a solution to classifying cyber biosensor systems. The algorithm of scientific machine learning includes the following stages: the definition of system trajectories classes; the generation of a matrix of random initial values of time delay and speed parameters; the model implementation and classification of input data; the construction of a matrix of connections between initial values, the time delay and between speed parameters; the application of the induction tree of the C5.0 algorithm to the relationship between the initial values and velocity parameters.

The method allows to optimally select the parameters of the cyber-physical biosensor system (initial conditions, speed parameters, and immune response time).

The main scientific results of the thesis were published in 11 works, particularly: two articles [1], [3] in scientific professional periodicals of Ukraine; one article in a foreign professional periodical [2]; one patent for a utility model was issued [11]; 7 publications [4], [5], [6], [7], [8], [9], [10] in the materials of international and pan-Ukrainian scientific and technical conferences. One work is part of the international scientometric database Scopus [4], two works [1], [3] are part of the international scientometric database Index Copernicus.

Keywords: machine learning, decision making under uncertainties, systematic medical research, deep learning, scientific machine learning, minimax approach, decision tree, cloud computing.