

ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

(повне найменування вищого навчального закладу)

Факультет математики, фізики та інформаційних технологій

(повне найменування інституту, назва факультету (відділення))

Кафедра математичного забезпечення комп'ютерних систем

(повна назва кафедри (предметної, циклової комісії))

Дипломна робота

на здобуття ступеня вищої освіти «магістр»

(освітньо-кваліфікаційний рівень)

на тему: Обробка великих масивів даних із застосуванням методів

машинного навчання

Processing of large data sets using machine learning methods

Виконав: студент денної форми навчання

спеціальності 123 – Комп'ютерна інженерія

(шифр і назва напряму підготовки, спеціальності)

Ярошук Олександр Валерійович

(прізвище, ім'я, по-батькові)

Керівник к.фіз.-мат.н., доц.Шпінарева І. М.

(науковий ступінь, вчене звання, прізвище та ініціали, підпис)

Рецензент д.техн.н., проф. Малахов Є.В.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент к.техн.н., доц. Рудніченко М.Д.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:

Захищено на засіданні ЕК № ____

Протокол засідання кафедри

протокол № __ від «__» ____ 2020 р.

№ ____ від «__» ____ 2020 р.

Оцінка ____ / ____ / ____

(за національною шкалою, шкалою ECTS, бали)

Завідувач кафедри

Голова ЕК

(підпис)

Є.В. Малахов
(прізвище, ініціали)

(підпис)

Н.Ф. Казакова
(прізвище, ініціали)

АНОТАЦІЯ

У дипломній роботі розробляється тема «Обробка великих масивів даних із застосуванням методів машинного навчання».

Аналіз і порівняння інформації є однією з найважливіших задач в будь-якій галузі людської діяльності, в тому числі в наукових і медичних дослідженнях. Обсяг записуваних в світових сховища даних щомиті зростає, а це означає, що такими ж темпами повинні змінюватися умови зберігання і обробка інформації. Під великими даними BigData зазвичай мають на увазі узагальнені набори, що включають в себе структуровані і неструктуровані дані, суттєві за обсягом і різноманітні за структурою. С зростанням обсягів даних набирають обертів і напрямки, в яких використовується обробка і аналіз великих даних. Більшість лікарень сьогодні використовують свого роду лікарняні інформаційні системи для управління своїми медичними даними чи даними пацієнтів. Ці системи зазвичай генерують величезні обсяги даних у вигляді чисел, тексту, діаграм і зображень. На жаль, ці дані рідка використовуються для підтримки прийняття клінічних рішень.

Метою роботи є підвищення точності прогнозування захворювання серця шляхом дослідження і модифікації методу дерева рішень.

В результаті проведених в роботі досліджень розглянути методи обробки великого обсягу медичних даних, досліджені алгоритми дерев рішень. У дипломній роботі сформовані вимоги до розробленої системи та запропонований модифікований алгоритм на основі алгоритму Gradient Boosting, що дозволило збільшити точність та зменшити складність. Виконано реалізація програми для прогнозування захворювання серця. В роботі проведено аналіз модифікованого алгоритму та алгоритмів CART, Random Forest, Gradient Boosting, ID3. Аналіз продемонстрував, що модифікований алгоритм має більшу точність.

В результаті проведеної роботи отримано алгоритм, який дозволяє більш точно прогнозувати захворювання серця.

АННОТАЦИЯ

В дипломной работе разрабатывается тема «Обработка больших массивов данных с применением методов машинного обучения».

Анализ и сравнение информации является одной из важнейших задач в любой области человеческой деятельности, в том числе в научных и медицинских исследованиях. Объем записываемых в мировых хранилища данных ежесекундно растет, а это значит, что такими же темпами должны меняться условия хранения и обработка информации. Под большими данными BigData обычно подразумевают обобщенные наборы, которые включают в себя структурированные и неструктурированные данные, существенные по объему и разнообразные по структуре. С ростом объемов данных набирают обороты и направления, в которых используется обработка и анализ больших данных. Большинство больниц сегодня используют своего рода больничные информационные системы для управления своими медицинскими данными или данными пациентов. Эти системы обычно генерируют огромные объемы данных в виде чисел, текста, диаграмм и изображений. К сожалению, эти данные редко используются для поддержки принятия клинических решений.

Целью работы является повышения точности прогнозирования заболевания сердца путем исследования и модификации метода дерева решений.

В результате проведенных в работе исследований рассмотрены методы обработки большого объема медицинских данных, исследованы алгоритмы деревьев решений. В дипломной работе сформированы требования к разработанной системе и предложен модифицированный алгоритм на основе алгоритма Gradient Boosting, что позволило увеличить точность и уменьшить сложность. Выполнена реализация программы для прогнозирования заболевания сердца. В работе проведен анализ модифицированного алгоритма и алгоритмов CART, Random Forest, Gradient Boosting, ID3. Анализ показал, что модифицированный алгоритм имеет большую точность. В результате проведенной работы получен алгоритм, который позволяет более точно прогнозировать заболевания сердца.

ABSTRACT

In this Master's Degree paper, the topic "Processing of large data sets using machine learning methods" is being revealed.

Analysis and comparison of information is one of the most important tasks in any area of human activity, including scientific and medical research. The volume of data recorded in the world's data storages is growing every second, which means that the conditions for storing and processing information must change at the same rate. Big data is usually understood as a generalized dataset, which includes structured and unstructured data that are significant in volume and varied in structure. With the growth of data volumes, the directions in which the processing and analysis of big data are used are gaining momentum.

Today most hospitals use some sort of hospital information system to manage their health or patient data. These systems typically generate huge data amounts in the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making.

The aim of this work is to improve the accuracy of predicting heart disease by researching and modifying the decision tree method.

As a result of the research carried out in this work, methods for processing a large amount of medical data were considered, and the algorithms of decision trees were investigated. In the thesis, the requirements for the developed system were formed and a modified algorithm based on the Gradient Boosting algorithm was proposed, which made it possible to increase the accuracy and reduce the complexity. The software implementation of the program for predicting heart disease is completed. The modified algorithm and algorithms CART, Random Forest, Gradient Boosting, ID3 are analyzed. The analysis showed that the modified algorithm is more accurate.

As a result of the work, an algorithm was obtained which allows to more accurately predict heart diseases.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ ТА ТЕРМІНІВ	6
ВСТУП	7
1 ДОСЛІДЖЕННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ І ОГЛЯД ПОПУЛЯРНИХ РІШЕНЬ.....	9
1.1 Огляд популярних систем, які використовують BigData	9
1.2 Класифікація методів машинного навчання.....	10
1.3 Класифікація методів прогнозування.....	11
1.4 Огляд існуючих і відповідних моделей часових рядів	13
1.4.1 Нейромережеві моделі	13
1.4.2 Моделі на базі ланцюгів Маркова	14
1.4.3 Моделі на базі класифікаційно-регресійних дерев ID3, C4.5, CART, SLIQ, RF, Gradient Boostig.....	15
1.5 Висновки та уточнення задач	24
2 МОДИФІКАЦІЯ АЛГОРИТМУ GRADIENT BOOSTING.....	26
2.1 Проектування системи	26
2.2 Перетворення та очистки вхідних даних	26
2.3 Алгоритм побудови дерева.....	29
2.4 Модифікації алгоритму Gradient Boosting	32
2.5 Засоби реалізації	32
3 РЕАЛІЗАЦІЯ СИСТЕМИ.....	34
3.1 Опис програмної реалізації	34
3.1.1 Модуль інтерпретації вхідних даних	34
3.1.2 Модуль побудови дерева.....	35
3.1.3 Модуль тестування.....	36
3.2 Інструкції для роботи з програмою	36
4 ОЦІНКА ЕФЕКТИВНОСТІ МОДИФІКОВАНОГО АЛГОРИТМУ	38
ВИСНОВКИ.....	45
СПИСОК ЛІТЕРАТУРИ.....	46
ДОДАТОК А Код застосунка	48

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ ТА ТЕРМІНІВ

CART – Classification And Regression Tree – Дерево Класифікації і Регресії'

GB – Gradient Boosting

RF – Random forest – випадковий ліс

SLIQ – Supervised Learning In QUEST

ІНМ – штучна нейронна мережа

ТЕРМІН

Ансамблювання – побудови алгоритму машинного навчання на базі кількох, в даному випадку, вирішальних дерев.

ВСТУП

Аналіз і порівняння інформації є однією з найважливіших задач в будь-якій галузі людської діяльності, в тому числі в наукових і медичних дослідженнях. Обсяг записуваних в світових сховища даних щомиті зростає, а це означає, що такими ж темпами повинні змінюватися умови зберігання і обробка інформації. Згідно з дослідженням IDC Digital Universe, в найближчі п'ять років обсяг даних на планеті зросте до 175 зеттабайт [1]. Під великими даними BigData зазвичай мають на увазі узагальнені набори, що включають в себе структуровані і неструктуровані дані, суттєві за обсягом і різноманітні за структурою. С зростанням обсягів даних набирають обертів і напрямки, в яких використовується обробка і аналіз великих даних

Більшість лікарень сьогодні використовують свого роду лікарняні інформаційні системи для управління своїми медичними даними чи даними пацієнтів. Ці системи зазвичай генерують величезні обсяги даних у вигляді чисел, тексту, діаграм і зображень. На жаль, ці дані зрідка використовуються для підтримки прийняття клінічних рішень. Аналіз великих даних веде до нової медицини – персоналізованої і превентивної. А значну частину рутини в роботі лікаря можна автоматизувати за рахунок попереднього збору і первинного аналізу даних. Лікар повинен приймати рішення, «перетравивши» вступну інформацію, але найчастіше її так багато, що пацієнт відсікає епізоди, які можуть мати вирішальне значення. В Україні в сфері медицини особливо виділяються серцево-судинні захворювання. За статистикою, відсоток хворих на туберкульоз різновидом цих захворювань перевищують середній відсоток в Європі більш ніж в 2 рази.

Для роботи з BigData застосовують методи машинного навчання. У 1980-х роках однією з центральних парадигм штучного інтелекту стало машинне навчання. Машинне навчання – дуже великий напрямок досліджень, перекривається з багатьма іншими областями і включає розгляд різних проблем на основі різних підходів, для яких навіть немає усталеної класифікації.

Основні задачі, які розв'язуються за допомогою методів машинного навчання:

- 1) класифікація – виконується за допомогою навчання з учителем;
- 2) регресія – виконується за допомогою навчання з учителем на етапі тестування, є окремим випадком задач прогнозування;
- 3) кластеризація – виконується за допомогою навчання без вчителя.

Одним з напрямків машинного навчання, що активно розвиваються, є розробка алгоритмів дерев рішень. Дерево рішень – спосіб прийняття рішення, заснований на застосуванні різних функцій поділу вихідного набору даних, зокрема простих порогових правил [2]. Деревом рішень називається бінарне дерево, в якому кожній внутрішній вершині приписана функція, а кожній листовій вершині приписаний прогноз.

У даній роботі пропонується аналізувати великий обсяг медичних даних з метою прогнозування захворювання серця.

Метою даної роботи є підвищення точності прогнозування захворювання серця шляхом дослідження і модифікації методу дерева рішень.

Об'єктом дослідження є процес інтелектуального аналізу даних в задачах прогнозування діагностики серцево-судинних захворювань.

Предметом дослідження виступають методи і алгоритми обробки великих масивів даних.

Для досягнення поставленої мети необхідно розв'язати такі задачі:

- дослідження методів дерев рішень для задач прогнозування;
- реалізація та аналіз методів дерев;
- проектування системи для прогнозування серцевих захворювань;
- реалізація обраного методу дерева рішень та його модифікації ;
- реалізація системи прогнозування в конкретній предметній області;
- аналіз модифікованого алгоритму.

ВИСНОВКИ

В рамках даної дипломної роботи досліджені існуючі методи машинного навчання. Досліджені алгоритми побудови дерев рішень: CART, Random Forest, ID3, C4.5. Дані алгоритми протестовані на базі Heart Diseases і дали точність від 75,5 до 89,6%. Обрано найкраще підходящий алгоритм RF. Ним побудовано дерево та навчено за допомогою алгоритму Gradient Boosting для прогнозування серцево-судинних захворювань.

Проведено проектування застосунку на обраному інструментальному середовищі розробки та виконана програмна реалізація системи для прогнозування захворювання серця.

В роботі виконана модифікація алгоритму Gradient Boosting. Попередня обробка даних дала результат на стандартному алгоритмі Gradient Boosting та алгоритмі з використанням випадкових підпросторів.

Модифікація алгоритму Gradient Boosting, в якій видалявся шум, не виправдала себе, оскільки точність була нижчою, ніж у алгоритмі з використанням випадкових підпросторів, та лише трохи вищою, ніж у базовому алгоритмі.

Модифікація алгоритму випадкових підпросторів допомогла значно збільшити точність алгоритму, та разом з попередньою обробкою даних цей алгоритм дав дуже хорошу точність.

Отже, можна зробити висновок, що модифікація алгоритму з використанням випадкових підпросторів є цілком виправдана та позитивна.

В роботі проведений аналіз модифікованого алгоритму та алгоритмів CART, Random Forest, Gradient Boosting, ID3. Аналіз продемонстрував, що модифікований алгоритм має більшу точність – 94%.

В результаті проведеної роботи отримано алгоритм, який дозволяє більш точно прогнозувати захворювання серця в порівнянні з алгоритмами CART, Random Forest, ID3, Gradient Boosting.

СПИСОК ЛІТЕРАТУРИ

1. IDC: Expect 175 zettabytes of data worldwide by 2025 [Електронний ресурс]. – Режим доступу: <https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html>
2. Awake Security Platform [Електронний ресурс]. – Режим доступу: <https://towardsdatascience.com/decision-tree-22975f00f3e1/>
3. Awake Security Platform [Електронний ресурс]. – Режим доступу: <https://awakesecurity.com/datasheets/awake-security-platform-datasheet/>
4. SAS Anti-Money Laundering [Електронний ресурс]. – Режим доступу: https://www.sas.com/ru_ua/software/anti-money-laundering.html
5. CHEQ [Електронний ресурс]. – Режим доступу: https://www.cheq.ai/about_us
6. How Netflix Uses AI, Data Science, and Machine Learning — From A Product Perspective [Електронний ресурс]. – Режим доступу: <https://becominghuman.ai/how-netflix-uses-ai-and-machine-learning-a087614630fe>
7. Jyoti. S, Uzma. A, Dipesh. S, Sunita. S. Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. International Journal on Computer Science Engineering, 2011; 3(6): 2385-2392.
8. Ramprasanth, H.; Devi, A. Outlier Analysis of Medical Dataset Using Clustering Algorithms. J. Anal. Comput. 2019, 15, 1–9.
9. Kodali. L, Adusumilli A.S, Doreti. P, Renuga Devi. T, Umamakeswari. A. Performance Analysis of Various Data Mining Techniques in the Prediction of Heart Disease. Indian Journal of Science and Technology, 2015; 8(35): 1-7.
10. Штучна нейронна сітка [Електронний ресурс]/ Режим доступу: http://www.machinelearning.ru/wiki/index.php?title=Искусственная_нейронная_сеть
11. Ланцюги Маркова [Електронний ресурс]/ Режим доступу: <https://towardsdatascience.com/introduction-to-markov-chains-50da3645a50d>

12. CART [Електронний ресурс]/ Режим доступа: <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/>
13. Оцінка класифікатору (точність, повнота, F-мера) [Електронний ресурс]. – Режим доступу: <http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html>
14. ID3 [Електронний ресурс]/ Режим доступа: <https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1>
15. C4.5 [Електронний ресурс]/ Режим доступа: <https://towardsdatascience.com/what-is-the-c4-5-algorithm-and-how-does-it-work-2b971a9e7db0>
16. R. Agrawal et. al. An interval classifier for database mining applications. In Proc. of the VLDB Conf., Vancouver, British Columbia, Canada, August 1992.
17. Gradient Boosting [Електронний ресурс]/ Режим доступа: <https://machinelearningmastery.com/gradient-boosting-algorithm-machine-learning/>
18. Yaroshchuk O.V. Decision trees in forecasting problems/ Yaroshchuk O.V., Yakushyna A.O., Ph.D. Shpinareva I.M.// Матеріали 9-ї міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології (ІУСТ–2020)» 24–26 вересня 2020 р. – Одеса, ОНПУ, 2020 – С.136-138
19. Heart Disease Data Set [Електронний ресурс]. – Режим доступу: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>.
20. Pandas [Електронний ресурс]. – Режим доступу: <https://pandas.pydata.org/>
21. Matplotlib [Електронний ресурс]. – Режим доступу: <https://matplotlib.org/>
22. Ярощук О. В., Якушина А. О., Шпінарева І. М. Порівняння методів побудови дерев рішень в обробці великих масивів даних// Інформатика, інформаційні системи та технології: XVII Всеукр. конференції студентів і молодих науковців. Одеса, 2020 р. – С.150-152
23. Suganya. R, Rajaram. S, Sheik Abdullah. A, Rajendran. V. A Novel Feature Selection Method for Predicting Heart Diseases with Data Mining Techniques. Asian Journal of Information Technology. 2016; 15(8): 1314-1321.